Teorijske raspodele verovatnoće

Slučajno promenljiva nastaje kada se određenim slučajnim događajima dodeljuju brojevi ili atributi, što zavisi od skale merenja, tako je da uz svaki dodeljeni broj ili atribut vezana neka verovatnoća.

Dakle, neka varijabla ***x*** može poprimiti ***k*** različitih vrednosti (npr. vrednost glikemije može biti 3,4; 5,6; 9,9mmol/l, ... do k). Ova varijabla postaje slučajno promenljiva ukoliko dobija pojedine vrednosti sa određenom verovatnoćom (npr. verovatnoća glikemije 5,6 u nekoj populaciji je 0,04). Raspodelu verovatnoće neke slučajno promenljive čine sve vrednosti koje ta varijabla može imati, kao i njene pripadajuće verovatnoće. Na taj način se formiraju krivulje verovatnoće. U praksi najčešće ove krivulje nisu pravilnog oblika. Ukoliko uspemo da aproksimiramo neke empirijske vrednosti u neku teorijsku raspodelu, u mogućnosti smo da vršimo najrazličitije procene i analize bioloških pojava. Značaj primene teorijskih rasporeda se ogleda u tome jer su kod njih proučeni svi elementi i sve osobine, a kod mnogih posmatranih rasporeda nisu uvek poznati svi elementi.

Teorijske raspodele verovatnoće prikazane su različitim matematičkim modelima. One prikazuju način na koji je ukupna verovatnoća raspodeljena na pojedine vrednosti slučajno promenljive. Zavisno od tipa slučajno promenljive teorijske rasodele verovatnoće dele se na: **prekidne** i **neprekidne**.

Raspodele prekidne slučajno promenljive prikazuju se grafički pologonom i štapićastim dijagramom, a raspodele neprekidne slučajno promenljive predstavljaju se krivom koja se naziva kriva gustine verovatnoće. Raspodele verovatnoće opisuje se najčešće određivanjem najmanje dva deskriptivna parametra (npr. aritmetička sredina, standardna devijacija, varijanasa itd.).

Najpoznatije teorijske raspodele su: binomni raspored, normalan raspored i Paissonov raspored. Osim ove tri teorijske raspodele često se u praksi primenjuju i Studentov t-raspored, χ^2-raspored, Fisherov raspored.

Binomni raspored

Binomni raspored predstavlja tip prekidne teorijske raspodele. Prvi put je opisao švajcarski matematičar Jacob Bernoulli 1700. godine.

Teorijski binomni raspored predstavlja raspored verovatnoća vrednosti prekidne slučajno promenljive dihotomnog, odnosno binomnog karaktera.

Prekidna slučajno promenljiva može da se raspoređuje po ovom matematičkom modelu pod uslovom da ima dva međusobno isljučiva ishoda (npr. uspeh/neuspeh).

Binomni raspored se zasniva na teoriji verovatnoće i zakonu velikih brojeva uz uslov da je p=0,5 i q=0,5, odnosno da je verovatnoća povoljnog i nepovoljnog događaja ista, tj. 50% i 50%. Primer za ovaj tip rasporeda je bacanje novčića pri čemu mogu da padnu pismo ili glava. Pismo i glava imaju podjednake šanse da se padnu p=0,5 i q=0,5. Međutim, u praksi se retko dešava da se 10 bacanja padnu 5 glava i 5 pisma. Dakle, nismo u mogućnosti da predvidimo ishod događaja. Na ishod utiču razni sporedni faktori: način bacanja, sila kojom je novčić bačen uvis, sam oblik i sastav novčića itd. Ipak na velikom broju pokušaja ishodi se približavaju svojim teorijskim verovatnoćama i to je postulat teorije verovatnoće tj. Zakon velikih brojeva.

Uz pomoć matematičkog obrasca može se odredi verovatnoća očekivanog događaja koji će se desiti ***x*** puta iz ***n*** pokušaja. Matematički obrazac za verovatnoću nastanka ***x*** očekivanih događaja iz ***n*** pokušaja glasi:

$$p\ (x;n)=\binom{n}{x}p^x \cdot q^{n-x},\ \binom{n}{x}=\frac{n!}{x!(n-x)!}$$

Simbol $\binom{n}{x}$ predstavlja skraćenicu za binomni koeficijent.

U kome je znak uzvika oznaka za faktorijel. Podsetimo se da faktorijel, na primer broja 5 iznosi: 5!=5x4x3x2x1.

U praksi se često radi radi lakšeg određivanja mogućih kombinacija umesto binomnog obrasca, čije je izračunavanje složeno koristi tzv. Paskalov trougao.

Osnovni uslovi za primenu binomnog rasporeda:

- posmatrana slučajno promenljiva mora da bude binomnog karaktera
- pokušaji mora da budu međusobno nezavisni ili da uzorak bude slučajno odabran
- da svaki događaj ima istu verovatnoću uspeha pri svakom pokušaju ili da svi članovi posmatrane populacije imaju istu šansu da budu slučajno izabrani
- veličina uzorka ***n*** mora da bude zanemarljiva u odnosu na veličinu populacije ***N***

Binomni raspored je definisan pomoću dva parametra, n i p. Sa **n** se označava broj jedinki u uzorku ili broj ponavljanja. Drugi bitan parametar koji definiše ovaj raspored je **p** odnosno stvarna verovatnoća uspeha za svaku jedinku ili svaki pokušaj. Zato se binomni raspored simbolički predstavlja kao B(n,p).

Aritmetička sredina binomne distribucije je vrednost slučajne varijable koju očekujemo ako posmatramo n jedinki ili pokušaj ponavljamo **n** puta, i iznosi np, a varijansa binomne distribucije iznosi np(1-p).

Kada je **n** mala vrednost, distribucija je zakrivljena u desno ako je $p<0,5$ a ako je $p>0,5$ distribucija je zakrivljena u levo. Sa porastom veličine uzorka distribucija postaje više simetrična. Binomna distribucija aproksimira normalnu distribuciju ako su joj vrednosti i aritmetičke sredine i varjanse veće od 5.

Normalan raspored

Teorijska distribucija koja je našla najveću primenu u medicinskim istraživanjima je normalan raspored. Posebno pri zaključivanju o osnovnom skupu na osnovu uzorka. Pripada tipu neprekidnih teorijskih rasporeda, tj. raspodela verovatnoće kontinuirane slučajno promenljive.

Normalan raspored je prvi proučavao francuski naučnik Abraham de Moivre (1667-1745). Naziv je dobio po nemačkom naučniku Karlu Friedrichu Gaussu (1777-1855), koji je proučavao greške merenja.

Normalan raspored je karakterističan za numerička neprekidna obeležja. Grafički se normalan raspored prikazuje Gausovom krivom.

Osobine normalne krive:

1. Po svom obliku normalan raspored je predstavljen kontinuiranom krivom zvonastog oblika, koja je unimodalna.
2. Proteže se od +∞ do –∞, asimptotski se približava x – osi; tako da je interval krive konačan.
3. Simetrična je u odnosu na aritmetičku sredinu. Znači da se 50% većih vrednosti nalazi levo od $\bar{x}$, a 50% manjih vrednosti nalazi desno od $\bar{x}$. Oko aritmetičke sredine se koncetriše najveći broj vrednosti, a ekstremne mala i ekstremno velike vrednosti imaju najmanju frekvenciju.
4. Kako je kriva simetrična aritmetička sredina je jednaka medijani odnosno modu: $\bar{x}=M_e=M_o$.
5. Normalan raspored je definisan $\bar{x}$, SD. Za svaki niz za koji je određen $\bar{x}$, SD moguće je konstruisati krivu. Aritmetička sredina određuje mesto zvona u koordinatnom sistemu, a standardna devijacija određuje nagib krive, tj. tačke pregiba krive.
6. Površina ispod normalne krive koja je ograničena x-osom izjednačava se sa 1, odnosno sa 100%
7. Za ovaj raspored važe sigma pravila:
   - U intervalu od ±1SD u odnosu na aritmetičku sredinu nalazi se 68,27% svih vrednosti. Dakle u ovom intervalu se nalazi:

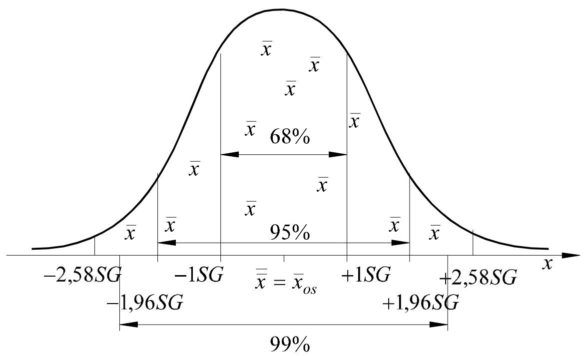
     $$P(\bar{x}-1SD<x>\bar{x}+1SD)=0,6827=68,27\%$$

   - U intervalu od ±2SD u odnosu na aritmetičku sredinu nalazi se 95,47% svih vrednosti. Dakle u ovom intervalu se nalazi:

     $$P(\bar{x}-2SD<x>\bar{x}+2SD)=0,9546=95,46\%$$

   - U intervalu od ±3SD u odnosu na aritmetičku sredinu nalazi se 99,73% svih vrednosti. Dakle u ovom intervalu se nalazi:

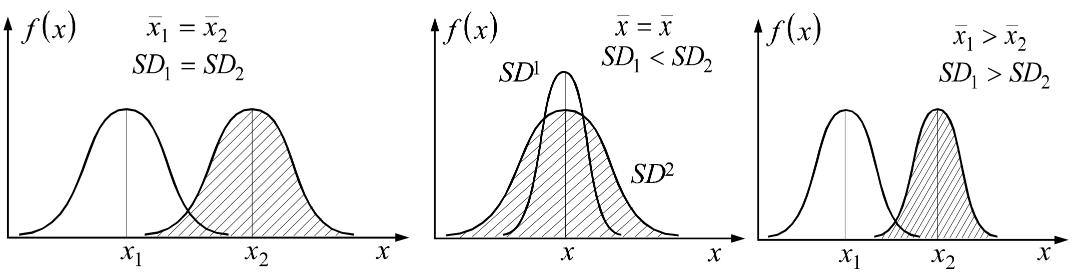
     $$P(\bar{x}-3SD<x>\bar{x}+3SD)=0,9973=99,73\%$$



Sve vrednosti slučajno promenljive nalaze se u okviru 6 standardnih devijacija, van ovog intervala ostaje zanemarljiv broj vrednosti (po 0,15% od intervala +∞ do –∞). Na osnovu ove osobine normalna raspodela može da se koristi kao aproksimacija svih ostalih rasporeda.

8. Ukoliko se poveća aritmetička sredina krivulja se pomera udesno, ukoliko se smanji pomera se ulevo. Ukoliko se poveća SD, krivulja se snižava i širi, ukoliko se smanjuje SD, krivulja je veća i uža. Slika ilustruje upravo oblike normalnog rasporeda definisane za različite parametre $\bar{x}$ i SD.

9. Ako slučajno promenljiva **x** ima normalan raspored, tada će i njena lienarna transformacija y=a+bx, takođe imati normalan raspored i suma **n** nezavisnih promenljivih ima normalan raspored.



Slika. Oblici normalnog rasporeda u zavisnosti od vrednosti aritmetičke sredine i standardne devijacije.

Standardizovan normalan raspored

Obzirom da je svaka Gausova kriva određena aritmetičkom sredinom i standardnom devijacijom, potrebno je da se za svaki niz podataka konstruiše kriva što zahteva dobro poznavanje matematičkih modela. Kako bi se upotreba ove krive učinila jednostavnijom definisan je standardizovan normalan raspored.

Standardizovan normalan raspored definisan je $\bar{x}=0$, SD=1. Osobine standardizovanog normalnog rasporeda iste su kao i kod normalnog rasporeda.

Ako podatke koji karakterišu standardizovan normalan raspored zamenimo u formuli za z – vrednost:

$$z = \frac{x - \bar{x}}{SD} = \frac{x - 0}{1} = x$$

dobićemo standardizovane vrednosti slučajno promenljive.

Na osnovu ovoga formirane su tablice verovatnoće odgovarajućih površina. U tabeli Površina ispod normalne kriva date su vrednosti funkcije rasporeda standardizovane normalne promenljive. U pretkoloni su date vrednosti z od 0 do 5. U središnjem delu tabele su date vrednosti površine ispod normalne krive između 0 i date z – vrednosti. Obzirom da je jedna od osobina ovog rasporeda simetričnost, jednake su površine između 0 i z , kao i između -z i 0.

Ako želimo da odredimo kolika je površina tj. verovatnoća za z=0,25 onda u prekoloni nađemo vrednost 0,2, a u zaglavlju vrednost 5 (označava drugu decimalu), na njihovom preseku očitamo vrednost 0987. Kako se verovatnoća označava ili od 0 do 1, ili u procentima. Očitanu vrednost možemo pisati kao 0,0987 ili 9,87%.

$z=\frac{x-\bar{x}}{SD}$	Druga decimala									
	0	1	2	3	4	5	6	7	8	9
0,0	0000	0040	0080	0120	0160	0199	0239	0279	0319	0359
0,1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0,2	0793	0932	0871	0910	0948	0987	1026	1064	1103	1141
0,3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...

U slučaju kada želimo da na osnovu verovatnoće odredimo z – vrednost onda je potrebno da datu verovatnoću podelimo sa 2, zbog osobine da su jednake površine između 0 i z , kao i između -z i 0. Tako za p= 40%, podelimo datu vrednost sa 2, i potražimo u tablici najbližu vrednost. Za p = 40% (p =0,4000) z =0,52.

Primer.

Pretpostavimo da je prosečna visina studenata medicine je 170, a SD=5, i da je raspoređena u vidu normalnog rasporeda.

$$\bar{x}=170, \text{ SD}=5.$$

a) Koji procenat studenata ima visinu od 170 do 180cm?

Obzirom da je jedna od vrednosti jednaka $\bar{x}$, onda je potrebno odrediti z – vrednost samo za 180cm

$$z = \frac{x - \bar{x}}{SD} = \frac{180 - 170}{5} = 2 \text{, za z = 2, p = 47,72\%}$$

Zaključujemo da 47,72% studenata ima visinu od 170 do 180cm.

b) Koji procenat studenata ima visinu od 162 do 172cm?ž

Potrebno je odrediti dve z-vrednosti, tj. za svaku visinu svoju z-vrednost.

$$h_1=162,\ z_1 = \frac{x-\bar{x}}{SD} = \frac{162-170}{5} = -1{,}6\text{, za } z = -1{,}6\ p_1=44{,}52\%$$

$$h_2=172,\ z_2 = \frac{x-\bar{x}}{SD} = \frac{172-170}{5} = 0{,}4\text{, za } z = 0{,}4,\ p_2 = 15{,}54\%$$

$$p = p_1 + p_2 = 44{,}52\% + 15{,}54\% = 60{,}06\%$$

c) Koji procenat studenata je niži od 160?

$$h_1=160,\ z_1 = \frac{x-\bar{x}}{SD} = \frac{160-170}{5} = -2\text{, za } z = -2,\ p_1=47{,}72\%$$

Dobijena verovatnoća pokazuje da se u intervalu od 160 – 170 nalazi 47,72% studenata, kako bi odredili procenat koji je niži od 160 potrebno je da oduzmemo određenu verovatnoću od 50%

$$p= 50\% - 47{,}72\% = 2{,}28\%$$

Studentov t - raspored

Studentov t – raspored je otkrio Vilijam Goset, hemičar i statističar zaposlen u Ginisovoj pivskoj kompaniji. Smatrao se studentom koji još uvek uči statistiku, tako da je svoje radove potpisivao pod pseudonimom „student“, ili je možda koristio pseudonim zbog Ginisovih ograničenja o tajnama zanata.

Kod malih uzoraka simbol z-vrednost se zamenjuje simbolom t-vrednost. Njegova formula je:

$$t=\frac{\bar{x}_{uz}-\bar{x}_{os}}{SG}=\frac{\bar{x}_{uz}-\bar{x}_{os}}{\frac{SD_{uz}}{\sqrt{n-1}}}$$

Suštinska razlika između t-vrednosti i z-vrednosti je u tome što u izrazu:

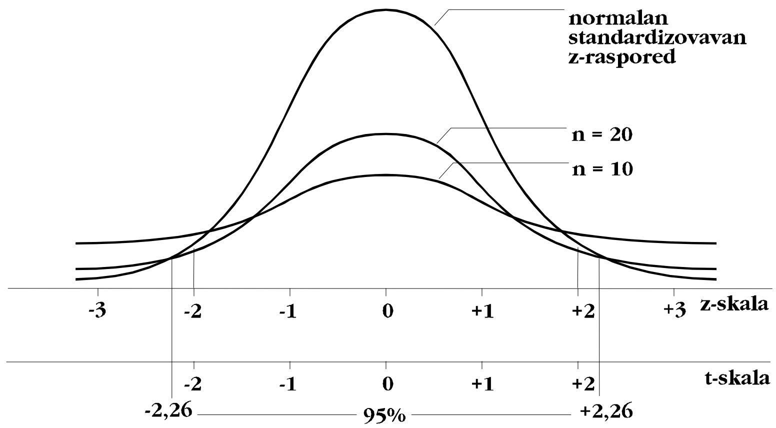
$$t=\frac{\bar{x}_{uz}-\bar{x}_{os}}{\frac{SD_{uz}}{\sqrt{n-1}}}$$

slučajno promenljive predstavljaju i $\bar{x}_{uz}-\bar{x}_{os}$ i $\frac{SD_{uz}}{\sqrt{n-1}}$.

U suštini standardna greška predstavlja ovde slučajno promenljivu, jer zavisi od veličine uzorka, samim tim t-vrednost zavisi od veličine uzorka, dok to nije bio slučaj sa z-vrednošću, jer je standardna greška bila konstantna, pa z-vrednost nije zavisila od veličine uzorka.

Aritmetičke sredine jednakih uzoraka, ali manjih od 30 jedinica raspoređuju se oko aritmetičke sredine osnovnog skupa, po posebnom t-rasporedu, tako da t-vrednost zavisi od veličine uzorka. Postoje različite vrste t distribucija, to je klasa distribucija.

Kriva t-rasporeda je simetrična i u obliku zvona kao kod normalnog rasporeda, unimodalna, definisana je od minus do plus beskonačno. Ipak, širenje je veće, dakle kriva je šira i spljoštenija u odnosu na normalni raspored. Kada se govori o određenoj t distribuciji, mora se odrediti stepen slobode. Sa porastom broja stepena slobode, t raspored bliži normalnom rasporedu.



Na slici je grafički prikazana razlika između t-rasporeda i standardizovanog normalnog rasporeda.

Za male uzorke za ocenu intervala ne mogu da se primene tablice normalnog rasporeda, već posebne Student-ove tablice t-rasporeda.

Širina krive uslovljava da se na nju ne može da primeni sigma tri pravilo, tj. da se u intervalu t= 0 ± 1,96 SG nalazi 95% t-vrednosti.

Ovaj procenat t-vrednosti se nalazi u znatno širem intervalu, a veličina intervala zavisi i od veličine uzorka: što je uzorak manji to je širina intervala veća.

- za uzorak od 2 jedinice u intervalu: $\bar{x}_{os} \pm 12{,}71 SG$
- za uzorak od 4 jedinice u intervalu: $\bar{x}_{os} \pm 3{,}18 SG$
- za uzorak od 10 jedinica: $\bar{x}_{os} \pm 2{,}26 SG$
- za uzorak od 20 jedinica: $\bar{x}_{os} \pm 2{,}09 SG$
- za uzorak od 30 jedinica: $\bar{x}_{os} \pm 2{,}04 SG$

Na osnovu iznetih osobina Studentovog t-rasporeda urađene su posebne tablice za granične t-vrednosti za odgovarajuću verovatnoću i broja stepena slobode.

Kod uzoraka većih od 30 jedinica i za odgovarajuću verovatnoću t-vrednost se izjednačuje sa z-vrednošću, a Studentov t-raspored zadobija oblik standardizovanog normalnog rasporeda.

Iz ovog se može zaključiti da tablice Studentovog t-rasporeda mogu da se primenjuju za ocenu intervala pouzdanosti aritmetičke sredine osnovnog skupa i na osnovu malih uzoraka ($n<30$ jedinica) i na osnovu velikih uzoraka ($n>30$). Odnosno da se t-raspored

mora koristiti pri radu sa malim uzorcima, a da je njegovo korišćenje preporučljivo pri radu sa velikim uzorcima.
U tabeli su prikazane osnovne razlike između z i t-vrednosti.

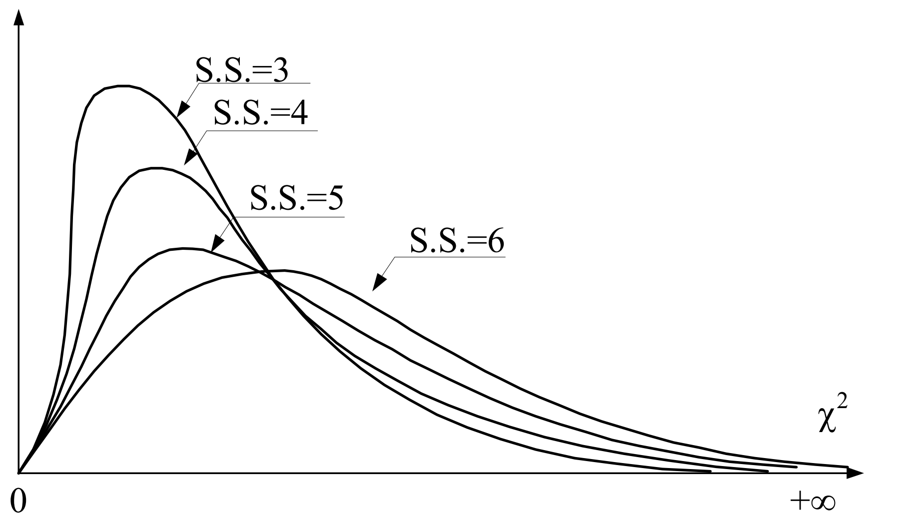
Razlike između Z i t – vrednosti

• Udaljenost slučajno promenljive od $\overline{X}$ izražena u jedinicama SD	• Odnos između razlike i SG te razlike
• Ako je $\overline{X}$ =0 a SD=1, Z se raspoređuje oko Z=0 u vidu standardizovanog normalnog rasporeda	• t-vrednost se oko t=0 ne raspoređuje u vidu standardizovanog normalnog rasporeda, već u vidu specifičnog t-rasporeda
• U intervalu Z=0 ± 3SD nalazi se 99,73% svih Z vrednost	• U intervalu t=0 ± 3SD nalazi se manje od 99,73% t-vrednosti.
• Za odgovarajuću P, Z se određuje na osnovu tablica normalnog rasporeda.	• Za odgovarajuću P i SS t se određuje na osnovu studentovih tablica t-rasporeda.
• Z=1,64 za P=0,90 • Z=1,96 za P=0.95 • Z=2,58 za P=0.99	• Sa povećanjem n, t se približava Z i sa n>30 izjednačava se sa Z.
• Testovi zasnovani na Z vrednosti primenjuju se samo za velike uzorke n>30.	• Testovi zasnovani na t-vrednosti mogu se primenjivati i na male i na velike uzorke.

Hi- kvadrat raspored

χ^2 distribucije je neprekidnog karaktera. Na osobinama χ^2 rasporeda bazira se zaključivanje prilikom korišćenja χ^2 testa. Tipičan grafik gustine ovog rasporeda prikazan je na slici. Osnove karakteristike Hi – kvadrat rasporeda:

- Raspored je definisan u oblasti od 0 do +∞;
- Kriva rasporeda nije simetrična, međutim sa povećanjem broja modaliteta posmatranog obeležja (sa povećanjem broja stepena slobode) Hi – kvadrat raspored se približava normalnom rasporedu.
- Za svaki stepen slobode postoji i određen Hi – kvadrat raspored i kritične oblasti prihvatanja ili odbacivanja nulte hipoteze.



Tumačenje realizovane vrednosti χ^2 testa vrši se na osnovu tablica kritičnih vrednosti χ^2 distribucije.

Zadaci za vežbanje

1. Kod 30 dijabetičara na klinici za endokrinologiju u Nišu određivan je nivo glikemije, dobijene su sledeće vrednosti:

H	F
6,2 – 7,49	**6**
7,5 – 8,79	**7**
8,8 – 10,09	**4**
10,1 – 11,39	**4**
11,4 – 12,69	**6**
12,7 – 13,99	**3**
Σ	**30**

   Sa verovatnoćom od 95% odredi kolika je prosečna glikemija kod svih dijabetičara u Nišu.

2. U niškom Kliničkom centru prosečno lečenje bolesnika je 50 dana ($\bar{x}$=50) i SD=5. Izračunati:
   - Koji procenat bolesnika je lečen manje od 25 dana?
   - Koji procenat bolesnika je lečen od 45 do 55 dana?
   - Koji procenat bolesnika je lečen više od 70 dana?
   - Koji procenat bolesnika je lečen od 20 do 30 dana?
   - Kako je niški klinički centar referentna zdravstvena ustanova u zemlji, proceniti kolika je prosečna dužina lečenja u svim kliničkim centrima u Srbiji sa verovatnoćom od 95%?