Lp 4 Rezumat

3.7. Distribuţia normală

Subiect de teoria probabilităţilor.

3.7.1. Distribuţie normală - descriere

Sinonime: *Distribuţie Gauss, distribuţie gaussiană, distribuţie Laplace, distribuţie laplaceană, distribuţie Gauss-Laplace, clopotul lui Gauss, curba erorilor (de măsurare întâmplătoare)*.
Descriere:

- Distribuţie continuă unimodală şi simetrică, cu cozi infinite care tind asimptotic la zero.
- Caracterizată de *media aritmetică* μ şi *abaterea standard* σ şi notată $N(\mu, \sigma)$.

3.7.2. Distribuţia normală standard şi consultarea tabelei corespunzătoare

- Este notată $N(0, 1)$.
- **Aria relativă α la dreapta punctului** z = -1,64 se citeşte pe linia –1,6 şi coloana -0,04. Se obţine aria α = 0,9495. (S-a dat *z=α-cuantila unilaterala superioară* şi s-a determinat α).
- **Punctul z care lasă la dreapta sa aria relativă α** = 0,95 este suma dintre valoarea liniei -1,6 şi a coloanei -0,04, la intersecţia cărora se află 0,9495, care este o cea mai apropiată valoare de α., ca şi 0,9505 aflat la intersecţia liniei –1,6 cu coloana –0,05. (S-a dat α şi s-a determinat *z=α-cuantila unilaterală superioară.*)

Determinarea altor arii şi *α-cuantilelor unilaterale inferioare*

- **Aria relativă la stânga** unui punct z = 1 - aria relativă la dreapta punctului z. (soluţia generală) = Aria la dreapta punctului $-z$. (Soluţie RECOMANDATĂ aici).
- **Aria relativă între z_1 şi z_2** (cu $z_1 < z_2$) = aria relativă la dreapta lui z_1 - aria relativă la dreapta lui z_2.

- **Punctul *z* care lasă la stânga sa aria relativă α =** punctul z care lasă la dreapta sa aria relativă 1- α (soluția generală) = punctul z care lasă la dreapta sa aria α, cu SEMN schimbat (soluție RECOMANDATĂ aici). (S-a dat α și s-a determinat *z = α-cuantila unilaterală inferioară.*)

3.7.3. Standardizare

- Scorul $z = (x - \mu) / \sigma$ transformă orice distribuție normală $N(\mu, \sigma)$ în cea standard, $N(0, 1)$, prin *centrare* și *reducere (standardizare).*
- Ariile relative legate de punctele *x* sub $N(\mu, \sigma)$ sunt egale cu ariile relative legate de punctele *z* sub $N(0, 1)$.

3.7.4. Corespondențe remarcabile

1° Arii pentru intervale sigmatice

Aria cuprinsă între **μ-3σ și μ+3σ** sub $N(\mu, \sigma)$ = aria cuprinsă între –3 și 3 sub $N(0,1)$ = **99,74 %.**
Aria cuprinsă între **μ-2σ** și **μ+2σ** sub $N(\mu, \sigma)$ = aria cuprinsă între –2 și 2 sub $N(0,1)$ = **95,44 %.**

2° α-cuantile unilaterale superioare remarcabile

$\alpha \cdot 100$ %:	10%	**5%**	**2,5%**	1%	0,5%	0,1%	0,05%
α-cuantila:	1,28	**1,64**	**1,96**	2,33	2,58	3,09	3,29

3.7.5. Aplicații în biologie

1° Scala de clasificare bazată pe intervale sigmatice [10]

Dimensiune:		**foarte mică**		**mică**		**medie**		**mare**		**foarte mare**	
Scală sigmatică	*M*-3*S*	→	*M*-2*S*	→	*M*-0,67*S*	→	*M*+0,67*S*	→	*M*+2*S*	→	*M*+3*S*

A fost concepută împreună cu scala centilică de la 3.4.2 (Lp2) pentru a coincide în cazul distribuirii gaussiene a datelor. (Vezi și http://app.inthlerom.ro/histo/HistoSetup.zip)

2° **Regula "trei sigma"** de eliminare a valorilor aberante : "Valorile din afara intervalului M-3S şi M+3S sunt eliminate din serie, *dacă datele se distribuie cvasigaussian, eventual printr-o transformare a lor şi au un volum mare, peste 30*."

3° Stabilirea **limitelor de normalitate biomedicală** (sau **regula "doi sigma"**) : "Sunt considerate normale valorile cuprinse între M-2S şi M+2S, *dacă datele se distribuie cvasigaussian, eventual printr-o transformare a lor şi au un volum cât mai mare.*"

3.7.6. Măsurarea gradului de concordanță cu o distribuție teoretică

în particular, distribuția normală (gaussiană).

- Prin $\chi^2 = \sum_{j=1}^{p} \frac{(O_j - T_j)^2}{T_j} = \sum_{j=1}^{p} \frac{O_j^2}{T_j} - N$ unde O_j sunt frecvențele observate (empirice) iar T_j sunt frecvențele teoretice în raport cu "teoria" stabilită: gaussianitate, uniformitate, etc.

Lp 4 Teste, exerciţii şi probleme

Următorul test grilă are acordat un timp mai mare deoarece este necesară consultarea tabelei cu α-cuantile superioare ale distribuţiei normale standard din Anexa 2.

TG4. Durata 250'' pe calculator.

Ce proportie α din aria de sub distributia normala standard se afla la dreapta punctului 1,87:
1. 0,0301
2. 0,0307
3. 0,9693

Ce proportie α din aria de sub distributia normala standard se afla la dreapta punctului 2,14:
1. 0,0125
2. 0,9838
3. 0,0162

Ce proportie α din aria de sub distributia normala standard se afla la dreapta punctului 2,63:
1. 0,0043
2. 0,0044
3. 0,9957

Valoarea -1,62 este α-cuantila superioara a distributiei normale standard cu α egal cu:
1. 0,9484
2. 0,0526
3. 0,9474

0,0099-cuantila inferioara a distributiei normale standard este:
1. 0,9901 cuantila superioara a distributiei normale standard
2. 0,0099 cuantila superioara a distributiei normale standard
3. 0,9901 cuantila inferioara a distributiei normale standard

Limitele de normalitate biomedicale se stabilesc:
1. aplicand regula 'doi sigma' datelor initiale
2. aplicand regula 'doi sigma' datelor eventual transformate pentru cvasigaussianizare
3. aplicand regula 'trei sigma' datelor eventual transformate pentru cvasigaussianizare

Distributia normala este caracterizata de urmatorii parametri:
1. media si abaterea standard
2. media si mediana
3. mediana, abaterea standard si coeficientul de variatie

Alegeti afirmatia ERONATA:
1. exista o infinitate de distributii normale
2. exista o singura distributie normala standard
3. exista o infinitate de distributii normale standard

α-cuantilele superioare pe o distributie normala oarecare se afla cu ajutorul tabelei pentru α cuantilele superioare ale distributiei normale standard dupa ce executam:
1. centrare
2. standardizare
3. reducere

TC4. Durata 3'.

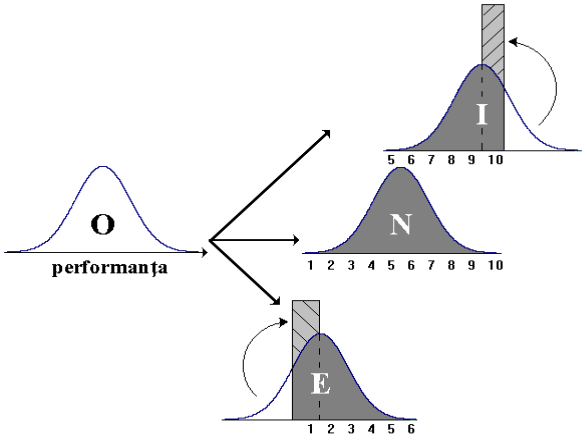
1. Distribuţia normală este o distribuţie continuă de forma unui _____ cu două cozi infinite ce tind _____ la zero.
2. Se numeşte standardizare a unei distribuţii normale aplicarea simultană a _____ şi _____ şi se obţine scorul z egal cu _____.
3. Dacă sub distribuţia normală standard aria la dreapta lui -3 este 0,9987, atunci aria la dreapta lui 3 este _____. Cât este aria cuprinsă între $\mu - 3\sigma$ şi $\mu + 3\sigma$ sub o distribuţie normală $N(\mu; \sigma)$? _____. Cât este aria la stânga lui $\mu + 3\sigma$? _____. Cât este aria la stânga lui $\mu - 3\sigma$? _____.

Exerciţii sau probleme rezolvate

1.

(a) Să se determine distribuţia notelor de la 1 la 10 pentru un lot omogen supus unui test de performanţă bine calibrat.
(b) Care este procentul de indivizi care nu promovează un astfel de test ?

Rezolvare:
(a)
Un lot omogen are performanţa distribuită gau-ssian (vezi distri-buţia "**O**" din figura alăturată). Pentru ca un test de apreciere a performanţei res-pective să fie bi-ne calibrat, va trebui ca notele de la 1 la 10 să fie acordate ca în cazul distribuţiei "**N**".



Dacă testul este prea **I**ndulgent, notele vor lua forma distribuţiei "**I**", care este extrem asimetrică de dreapta. Această formă provine din cumularea a două efecte. Primul este produs de translatarea distribuţiei către notele mari, ceea ce aduce un procent mult mai mare de note mari, în particular de 10. Se produce astfel distribuţia normală trunchiată la dreapta (vezi aria gri închis). Al doilea efect este creşterea şi mai exagerată a proporţiei de note de 10 din cauza acordării notei maxime tuturor celor care, din cauza slabei exigenţe, ar obţine aprecieri superioare lui 10 (vezi coada albă din dreapta distribuţiei "**I**" care devine aria haşurată din cadrul acesteia, arie adăugată între 9,5 şi 10 peste distribuţia normală trunchiată).
Invers, dacă testul este prea **E**xigent, notele vor lua forma distribuţiei "**E**". Aceasta este extrem asimetrică de stânga din motive analoage.
În concluzie, un test bine calibrat va produce o distribuţie a notelor cvasi**N**ormală, cum este distribuţia "**N**").
Pentru a calcula procentele fiecărei note, trebuie mai întâi să determinăm media şi abaterea standard ale distribuţiei gaussiene "**N**". Media trebuie să fie în centrul distribuţiei, deci $\mu = (1 + 10) / 2 = 5{,}5$. Abaterea standard, σ, o vom determina aplicând regula "doi sigma", considerând că nota 10 (care începe de

la 9,5) trebuie acordată celor care sunt mai performanți decât cei normali, respectiv, nota că nota 1 trebuie dată celor cu performanță inferioară normalilor. Cum nota 10 începe de la 9,5 rezultă că trebuie să plasăm valoarea 9,5 în limita superioară a normalității, $\mu + 2 \cdot \sigma$. Din egalitatea $\mu + 2 \cdot \sigma = 9{,}5$, adică $5{,}5 + 2 \cdot \sigma = 9{,}5$ rezultă $\sigma = 2$. (Pe baza simetriei distribuției normale rezultă «automat» că 1,5 se va plasa în limita inferioară a normalității, $\mu - 2 \cdot \sigma$.)

Pentru determinarea proporțiilor de arii corespunzătoare fiecărei note observăm că nota 1 se acordă prin rotunjire oricărui rezultat < 1,5, nota 2 înseamnă orice rezultat cuprins în intervalul [1,5; 2,5), etc. Nota 10 înseamnă un rezultat ≥ 9,5. În continuare, trebuie determinate scorurile $z_i = (x_i - \mu) / \sigma$ pentru limitele intervalelor respective. Apoi pentru fiecare scor z_i vom consulta tabela de α-cuantile superioare ale distribuției normale standard (vezi Anexa 2).

De exemplu, pentru nota 10 ne interesează proporția de arie aflată la dreapta lui 9,5, proporție notată $p(x \geq 9{,}5)$. Luând $\boldsymbol{x = 9{,}5} \Rightarrow z = (9{,}5 - 5{,}5) / 2 = 2$, adică $p(x \geq 9{,}5) = p(z \geq 2)$. Consultând tabela, pe linia 2 şi coloana 0, găsim $\boldsymbol{p(z \geq 2) = 0{,}0228}$.

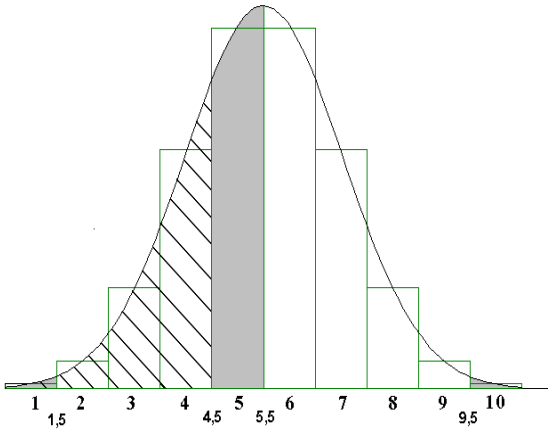
Tabela următoare conține marginile x_i ale claselor, scorurile z_i, proporțiile ariilor care se află la dreapta acestor puncte sub distribuția normală standard, precum şi proporțiile ariilor fiecărei clase.

limitele intervalelor, x_i	***scorurile z_i***	***$p(x \geq x_i) = p(z \geq z_i)$***	***$p(x_{i-1} \leq x < x_i) = p(z_{i-1} \leq z < z_i)$***		
-∞	-∞	1			
			0,0227	2 %	
1,5	-2	0,9773			
			0,0441	5 %	*
2,5	-1,5	0,9332			
			0,0919	9 %	
3,5	-1	0,8413			
			0,1498	15 %	
4,5	-0,5	0,6915			
			0,1915	19 %	
5,5	0	0,5000			
			0,1915	19 %	
6,5	0,5	0,3085			
			0,1498	15 %	
7,5	1	0,1587			

			0,0919	9 %	
8,5	1,5	0,0668			
			0,0440	5 %	*
9,5	**2**	**0,0228**			
			0,0228	2 %	
∞	-∞	0			

Ariile relative corespunzătoare fiecărei note se vor determina astfel:

- pentru nota 10 → $p\,(x \geq 9{,}5)$, adică proporţia de arie aflată la dreapta lui 9,5 = 0,0228.
- pentru notele 2, 3,..., 9 proporţia va fi diferenţa între proporţia de arie aflată la dreapta limitei inferioare şi proporţia de arie aflată la dreapta limitei superioare. De exemplu, pentru nota **5** (vezi figura alăturată), proporţia = proporţia de arie aflată la dreapta lui 4**,5** - proporţia de arie aflată la dreapta lui 5**,5** = $p\,(x \geq 4{,}5) - p\,(x \geq 5{,}5)$ = 0,6915 - 0,5 = **0,1915**.



- pentru nota 1 → $p\,(x < 1{,}5)$, adică proporţia de arie existentă la stânga lui **1,5** = 1 - $p\,(x \geq 1{,}5)$ = 1 - proporţia de arie existentă la dreapta lui 1,5 = 1 - 0,9773 = 0,0227.

Notă:

Valorile procentelor marcate prin semnul "*" au fost crescute puţin pentru ca totalul procentelor să fie 100 %.

(b)

Indivizii care nu promovează primesc note < 4,5 (vezi aria haşurată. Notele mai mari decât 4,5 dar mai mici decât 5 sunt evident rotunjite la 5.) Proporţia celor care nu promovează un test bine calibrat, $p\,(x < 4{,}5)$, va fi suma dintre proporţiile notelor de la 1 la 4 (adică toate rezultatele < 4,5) = $p\,(x < 1{,}5) + p\,(1{,}5 \leq x < 2{,}5) + p\,(2{,}5 \leq x < 3{,}5) + p\,(3{,}5 \leq x < 4{,}5)$ = 0,0227 + 0,0441 + 0,0919 + 0,1498 = 0,3085. Deci, procentul de nepromovaţi va fi 30,85% ≈ 31% ≈ 1/3, adică, aproximativ o persoană din trei nu va promova un test bine calibrat.

2.

S-a determinat nivelul limfocitelor T helper prin rozetare E cantitativă la un lot de 100 000 subiecţi consideraţi normali şi s-a observat o bună concordanţă între histograma distribuţiei empirice şi o curbă normală, ambele având media 45 şi abaterea standard 2,5. Uitând distribuţia empirică, să se determine:
(a) Proporţia şi numărul subiecţilor cu nivelul limfocitelor T helper ≥ 47.
(b) Proporţia şi numărul subiecţilor cu nivelul limfocitelor T helper < 47.
(c) Proporţia şi numărul subiecţilor cu nivelul limfocitelor T helper cuprins între 45 şi 47.
(d) Limitele de normalitate ale nivelului limfocitelor T helper.
(e) Diagnosticul de normalitate pentru un pacient cu nivelul limfocitelor T helper de 51.

Rezolvare:
(a) $z_a = (47 - 45) / 2,5 = 0,8$; $p\,(x \geq 47) = p\,(z \geq 0,8) = 0,2119$; $N_a = 0,2119 \cdot 100000 = 21190$.
(b) $p\,(x < 47) = p\,(z < 0,8) = 1 - p\,(z \geq 0,8) = 1 - 0,2119 = 0,7881$; $N_b = 0,7881 \cdot 100000 = 78810$.
(c) $p\,(45 \leq x < 47) = p\,(x \geq 45) - p\,(x \geq 47) = p\,(z \geq 0) - p\,(z \geq 0,8) = 0,5 - 0,2119 = 0,2881$; $N_c = 0,2881 \cdot 100000 = 28810$.
(d) Limitele de normalitate corespund valorilor $M - 2 \cdot S$ respectiv $M + 2 \cdot S$:
$M - 2 \cdot S = 45 - 2 \cdot 2,5 = 40$; $M + 2 \cdot S = 45 + 2 \cdot 2,5 = 50$.
(e) Valoarea 51 este mai mare decât limita superioară a intervalului de normalitate. Nivelul limfocitelor T helper la pacientul considerat nu este normal, ci este mai mare.

3.

S-au măsurat 10 tensiuni intraoculare la 10 persoane fără probleme oftalmologice şi s-au obţinut valorile următoare, măsurate în mm coloană de mercur:

20, 16, 24, 22, 20, 18, 20, 20, 20,20.

1. În vederea calculării limitelor de normalitate să se excludă eventualele valori aberante.
2. Să se stabilească limitele de normalitate.
3. Să se construiască scala cu trei trepte: hipotensiune, normotensiune, hipertensiune.
4. Să se specifice care este procentul aproximativ de normali în populaţie.
5. Presupunând că, în practică, cei cu tensiuni intraoculare peste L = 23 sunt suspectaţi de glaucom să se calculeze care este procentul de indivizi normali care sunt diagnosticaţi greşit astfel.
6. Aceeaşi întrebare dacă L = 24,5.

7. Care sunt probabilitățile ca alegând la întâmplare un individ normal să i se pună diagnosticul de glaucom atunci când utilizăm limită 23, respectiv, limita 24,5 ?
8. Ce putem face ca să micşorăm şi mai mult această probabilitate ?

(Problema se va rezolva având acces la manual şi calculator. Se va lua în considerație şi simpla prezentare a paşilor necesari de urmat, fără a se face calculul. Această descriere de algoritmi se va puncta mai mult decât rezultatele numerice corecte.)

Rezolvare:

1. Pentru detectarea şi eliminarea eventualelor valori aberante, trebuie calculate media (M) şi abaterea standard (S) ale seriei. Apoi, se vor calcula limitele M-3S, M+3S. Valorile din afara intervalului (M-3S, M+3S) sunt considerate valori aberante.
   Calculăm media şi dispersia prin formulele de calcul rapid şi exact:
   Pentru aceasta, multe valori repetându-se, este convenabil să grupăm seria în distribuția de frecvențe din primele două coloane ale tabelei următoare:

x_j	N_j	$x_j N_j$	$x_j^2 N_j$
16	1	16	256
18	1	18	324
20	6	120	2400
22	1	22	484
24	1	24	576
	N=10	$T_1 = 200$	$T_2 = 4040$

$M=T_1/N=200/10=20$

$S^2=T_2/N - M^2=4040/10-20^2=$
$=404-400=4$

De unde abaterea standard

$S=\sqrt{S^2} = \sqrt{4} = 2$

   - Limitele pentru excludere valori aberante vor fi deci:
   M-3S = 20 – 3 * 2= 14
   M+3S = 20 + 3* 2 = 26
   Se observă că nu există valori aberante.
2. Limitele de normalitate sunt:
   M - 2S = 20 – 2 * 2= 16
   M +2S = 20 + 2* 2 = 24
3. Scala cerută va fi:

$x < 16$	hipo
$16 \le x < 24$	normo
$24 \le x$	hiper

4. cca. 95%
5. Calculăm scorul z pentru valoarea L = 23:
   z = (L-M)/S = (23-20)/2=1,5
   Consultând tabela din anexa 2 se obține valoarea 0, 0668, adică cca. 7%.
6. Analog, z=2,25, iar aria = 0,0122 adică cca. 1%.

7. 7%, respectiv, 1%.
8. Mărim şi mai mult limita L.

Exerciţii sau probleme propuse

4.

S-a determinat nivelul limfocitelor T supresoare prin rozetare E cantitativă la un lot de 100000 subiecţi consideraţi normali şi s-a observat o bună concordanţă între histograma distribuţiei empirice şi o curbă normală, ambele având media 20 şi abaterea standard 2,5. Uitând distribuţia empirică, să se determine:
(a) Proporţia şi numărul subiecţilor cu nivelul limfocitelor T supresoare ≥ 23.
(b) Proporţia şi numărul subiecţilor cu nivelul limfocitelor T supresoare < 23.
(c) Proporţia şi numărul subiecţilor cu nivelul limfocitelor T supresoare cuprins între 20 şi 23
(d) Limitele de normalitate ale nivelului limfocitelor T supresoare.
(e) Diagnosticul de normalitate pentru un pacient cu nivelul limfocitelor T supresoare de 16.

5.

S-a măsurat tensiunea intraoculară la 8 persoane cu glaucom şi s-au obţinut valorile următoare, măsurate în mm coloană de mercur:

24, 26, 28, 28, 30, 30, 32, 34.

1. Să se excludă eventualele valori aberante.
2. Să se calculeze care sunt limitele între care se plasează în jurul mediei, în toate cazurile cu glaucom, cca. 95% din indivizi.
3. Presupunând că, în practică, cei cu tensiuni intraoculare peste L = 23 sunt suspectaţi de glaucom să se calculeze care este în toate cazurile posibile (nu numai pe acest eşantion), procentul de indivizi cu glaucom, care sunt diagnosticaţi greşit astfel.
4. Aceeaşi întrebare dacă L = 24,5.
5. Care sunt probabilităţile ca alegând la întâmplare un individ cu glaucom să i se pună un diagnostic diferit de glaucom atunci când utilizăm limită 24,5, respectiv, limita 23 ?
6. Ce putem face ca să micşorăm mai mult această probabilitate ?