Caracteristici numerice ale variabilelor aleatoare. Momente de selecţie. Legea numerelor mari. Teoreme limită centrală

Ne propunem să prezentăm legăturile între constantele numerice caracteristice unei repartiţii probabilistice date şi cele corespunzătoare datelor de selecţie.

1. Caracteristici numerice ale variabilelor aleatoare

În lucrările de laborator anterioare precizarea repartiţiei unei variabile aleatoare se realiza prin intermediul funcţiei de repartiţie sau a densităţii de repartiţie. Adesea, manipularea acestor funcţii este dificilă; din acest motiv se preferă o descriere prin intermediul unor constante numerice numite momente. Pentru variabila aleatoare $\xi = \begin{pmatrix} x_i \\ p_i = P[\xi = x_i] \end{pmatrix}, i = \overline{1,n}$, se definesc momentele:

Variabila aleatoare ξ	Momente absolute de ordin k	Momente centrate de ordin k
Discretă	$M[\xi^k] = \sum_{i=1}^{n} x_i^k p_i$	$M[(\xi - M[\xi])^k] = \sum_{i=1}^{n} (x_i - M[\xi])^k p_i$
Continuă	$M[\xi^k] = \int_{-\infty}^{+\infty} x^k f(x)dx$	$M[(\xi - M[\xi])^k] = \int_{-\infty}^{+\infty} (x - M[\xi])^k f(x)dx$

Observaţii:
1. pentru $k = 1$ M[ξ] se numeşte media (speranţa matematică) a variabilei ξ;
2. pentru $k = 2$, momentul centrat de ordin doi $M[(\xi - M[\xi])^2] = \sigma^2 = D[\xi]$ se numeşte dispersia dispersia variabilei ξ, iar σ reprezintă abaterea medie pătratică a variabilei aleatoare ξ.

2. Momente de selecţie. Legea numerelor mari. Teoreme limită centrală

Practic, atât funcţiile de repartiţie cât şi momentele sunt folosite pentru a caracteriza diferite populaţii statistice (colectivităţi statistice). Populaţiile statistice nu pot fi investigate prin considerarea tuturor elementelor componente, ci doar prin intermediul unor "eşantioane statistice". Astfel, dată fiind o populaţie de N indivizi, în urma a n experimente independente se obţin rezultatele $x_1, x_2, ..., x_n$, numite variabile de selecţie. Pe baza acestor date poate fi definită distribuţia în frecvenţe a variabilei aleatoare $\begin{pmatrix} x_i \\ f_i \end{pmatrix}, i = \overline{1,n}$ unde $\sum_{i=1}^{n} f_i = 1$ şi momentele de selecţie:

- **momentul de selecţie de ordin r**: $M_r = \sum_{i=1}^{k} f_i x_i^r$
- **media de selecţie** $\bar{x} = \sum_{i=1}^{k} f_i x_i$
- **momentul centrat de selecţie de ordin r** $m_r = \sum_{i=1}^{k} f_i (x_i - \bar{x})^r$
- **dispersia de selecţie** $\bar{s}^2 = \sum_{i=1}^{k} f_i (x_i - \bar{x})^2$
- **abaterea medie pătratică** $\bar{s}$

Conform legii numerelor mari frecvenţele relative converg către probabilităţile de apariţie a aceloraşi valori $f_i \xrightarrow[n\to\infty]{c.p.} p_i$, iar momentele de selecţie sunt estimaţii ale momentelor probabilistice.

Exemplu: În cele ce urmează vom ilustra convergenţa frecvenţei relative către probabilitate, şi vom evidenţia diferenţa între momentele teoretice şi cele de selecţie. Pentru aceasta vom alege o variabilă uniform distribuită, care caracterizează aruncarea cu zarul. Se aruncă un zar, dacă apare o faţă cu număr impar de puncte, vom câştiga o sumă egală cu numărul de puncte de pe faţă, în caz contrar vom pierde o sumă egală cu numărul par de puncte ale feţei.
Mai întâi se simulează jocul:

```
% program ZARURI
% simulează un joc în care câştigăm suma = nr. puncte de pe faţa
%impară şi pierdem suma = nr. puncte de pe faţa pară apărută în  urma
 %unei aruncări
 nr_ar=input('Numar aruncari : ');
cistig=0; freq=zeros(1,6); aport_face=zeros(1,6);
for i=1:nr_ar
     face=fix(6*rand)+1;
    freq(face)=freq(face)+1;
    rest=face-2*fix(face/2);
    if rest~=0,
         aport_face(face)=face;
   else
       aport_face(face)=-face;
   end
   cistig=cistig+aport_face(face);
end
freq_rel=zeros(1,6);
 disp('  Fata     Cistig    Frecventa    Frecventa relativa ');
for j=1:6
freq_rel(j)=freq(j)/nr_ar;
m=[j aport_face(j) freq(j) freq_rel(j)] ;
disp(m);
end
disp('Cistig mediu ');
disp(sum(aport_face.*freq_rel));
disp(' ');
```

Programul tipăreşte frecvenţa absolută şi relativă cu care apare fiecare faţă a zarului. În final se va afişa câştigul mediu. Pentru diferite numere de aruncări, rezultatele vor apărea astfel:

Faţa	**Câştig**	**Frecv. abs.**	**Frecv. relativă**	**Faţa**	**Câştig**	**Frecv. abs.**	**Frecv. relativă**
1	1	17	0.17	1	1	1681	0.1681
2	-2	17	0.17	2	-2	1678	0.1678
3	3	16	0.16	3	3	1626	0.1626
4	-4	18	0.18	4	-4	1696	0.1696
5	5	16	0.16	5	5	1686	0.1686
6	-6	16	0.16	6	-6	1633	0.1633

Nr aruncări=100
Câştig mediu= -0.57;

Nr aruncări=10.000
Câştig mediu= -0.4949;

Prima rulare ne sugerează că jocul este defavorabil. Ne punem problema cât de defavorabil este în realitate? A doua rulare oferă o imagine mai exactă. Se observă că frecvenţa relativă a fiecăreia din cele 6 feţe este foarte apropiată de 1/6=0.1666 (corespunde interpretării frecvenţiale a probabilităţii), adică frecvenţele relative converg în probabilitate către probabilităţile de apariţie a valorilor măsurate când volumul selecţiei tinde la infinit (pentru un număr de 10000 de aruncări frecvenţă relativă se apropie de valoarea de 0.16666, pe când în cazul unui număr de 100 de aruncări diferenţă este mai mare).
De asemeni, se sugerează că pentru un număr foarte mare de aruncări, câştigul mediu ar trebui să fie:

M = 1(1/6) - 2(1/6) + 3(1/6) - 4(1/6) + 5(1/6) - 6(1/6) = -0.5

iar dispersia:

$D = (1 + 0.5)^2*(1/6) + (-2 + 0.5)^2*(1/6) + (3 + 0.5)^2*(1/6) + (-4 + 0.5)^2*(1/6) + (5 + 0.5)^2*(1/6) = 14.9$ ($\sigma = 3.87$)

Pentru a determina dispersia de selecţie vom folosi următorul tabel:

Câştig	(Câştig-câştig mediu)*Frecvenţa relativă
1	$(1+0.4949)^2 * 0.1681$
-2	$(-2+4949)^2 * 0.1678$
3	$(3+0.4949)^2 * 0.1626$
-4	$(-4+0.4949)^2 * 0.1696$
5	$(5+0.4949)^2 * 0.1686$
-6	$(-6+0.4949)^2 * 0.1633$

$S^2 = \sum_{i=1}^{6}$ (Câştig-câştig mediu)^2 *frecvenţa relativă=14.86 => S=3.85

Execiţiu 1: Determinaţi, prin simulare, media şi dispersia de selecţie a unei variabile aleatoare repartizate uniform în intervalul [0,1]; comparaţi cu valorile teoretice M = 0.5 şi D = 1/12.
Exerciţiu 2: Determinaţi prin simulare media şi dispersia de selecţie a unei variabile aleatoare repartizate exponenţial negativ; comparaţi cu valorile teoretice M=1/*l* şi D = 1/*l*.
Exerciţiu 3: O problema tipică legată de repartiţia exponenţial negativă este problema "cozilor de aşteptare". Presupunem sosirea clienţilor la o staţie PECO, deservită de un singur vânzător la momente aleatoare de timp. Fiecare client este servit imediat dacă nu este nimeni înaintea sa; în caz contrar va trebui să aştepte deservirea clienţilor din faţa sa. Se pune problema cât timp va trebui să aştepte fiecare client?
Se presupune că intervalele de timp între sosirile succesive ale clienţilor sunt variabilele aleatoare X_1, X_2, ..., X_n ce sunt independente şi identic exponenţial distribuite, cu funcţia de repartiţie: $F_x(t)=1-e^{-lt}$, respectiv $f_x(t)=le^{-lt}$. Presupunem, de asemeni că timpul de deservire pentru fiecare client este definit prin variabilele aleatoare Y_1, Y_2,..., Y_n independente, identic exponenţial distribuite, cu funcţiile de repartiţie: $F_y(t)=1-e^{-mt}$ şi $f_y(t)=me^{-mt}$. Ne aşteptăm ca timpul petrecut de un client la coadă să depindă de numărul de clienţi din faţă să şi de timpul de deservire al fiecăruia. Scrieţi un program ce simulează *N*(*t*)=numărul de clienţi la coadă la momentul *t*, pentru *t*=0:100 şi *l*=1 şi *m*=0.9; *l*=1 şi *m*=1; *l*=1 şi *m*=1.1.
Observaţie:
Se poate realiza simularea pornind de la programul BUS PARADOX din laboratorul 3. În reprezentarea grafică pentru l<*m*, *N*(*t*) rămâne de dimensiune rezonabilă; iar în caz contrar *N*(*t*) creşte în mod nelimitat.

Vom defini: *l* = rata sosirilor; *m* = rata deservirii şi intensitatea traficului: r = $\frac{rata\ sosirilor}{rata\ deserviri} = \frac{1}{m}$.

Pentru $r < 1$, definim lungimea cozii (numărul de clienţi) ca o variabilă aleatoare *Z*, cu media finită: $M(Z) = N$. Timpul petrecut la coadă este de asemeni o variabilă aleatoare *W*, cu media: $M(W) = T$. Astfel, putem conchide că atunci când un client ajunge la coadă, el se aşteaptă să găsească *N* persoane înaintea sa, iar în momentul când părăseşte coada, va lăsa lT persoane după el. La echilibru, aceste valori trebuie să fie egale: $N = lT$. Cunoscând media variabilei aleatoare $N(t)$, să se estimeze $T = N/l$, adică media timpului de aşteptare.

După cum s-a observat din exemplele anterioare, mediile de selecţie (statistice) tind să se apropie de mediile (sau momentele) teoretice (probabilistice).
O serie de teoreme arată condiţiile în care o variabilă cu o anumită repartiţie converge către repartiţia normală sau condiţiile în care efectul cumulat al acţiunii unor variabile aleatoare, independent de repartiţiile lor, converg către un fenomen repartizat normal (v. curs).
În particular fie un proces aleator alcătuit din experimente succesive de tip binomial şi x_i variabile aleatoare de selecţie asociate încercării i. Se cunoaşte că dacă variabila aleatoare x este repartizată binomial de parametri n şi p şi $S_n = x_1 + ... + x_n$ este numărul de succese la cele n experimentări, atunci pentru $a < b$:

$$\lim_{n \to \infty} P\left(a < \frac{S_n - np}{\sqrt{npq}} < b\right) = \frac{1}{\sqrt{2\pi}} \cdot \int_a^b e^{-\frac{x^2}{2}} dx$$

Astfel, dacă variabila aleatoare x are o distribuţie de medie $M[x]$ şi abatere medie pătratică $\sigma[x]=\sqrt{D[x]}$, atunci media de selecţie este o variabilă aleatoare, ce va tinde în probabilitate la o repartiţie normală cu media $M[x]$ şi dispersia $D[x]/n$.
Pentru repartiţia binomială se pot determina analitic: $M_x = np$ şi $D_x=npq$. Practic s-a observat că pentru: $np > 5$, $nq > 5$ (condiţii îndeplinite simultan), unde: $q=1-p$ = probabilitatea de eşec într-o singură experimentare, variabilă x va tinde la o repartiţie normală de parametri: $M_{normal} = np$ şi $D_{normal} = npq$, dacă n este suficient de mare.
Exerciţiu 4: Reprezentaţi densitatea de probabilitate binomială pentru $p = 0.25$, $k = 0, ..., n$ şi n=3; 10; 30; 50. Pentru fiecare din cele 4 grafice obţinute, reprezentaţi şi densitatea de probabilitate normală corespunzătoare. Poate fi aproximată repartiţia binomială cu o repartiţie normală? Pentru ce valori ale lui n?
Exerciţiu 5: Presupunem un experiment de tip Bernoulli, de aruncare a monezii. Să se determine, prin simulare, media şi dispersia de selecţie pentru n = 20; 30; 40; 60; 100 şi să se compare cu valorile teoretice $M=np$, $D=npq$. Dacă rezultatele în urma efectuării a n aruncări sunt:
$x_1, x_2, ..., x_n$, să se determine: $S_n = x_1 + x_2 + ...+ x_n$ *şi* $S_n^* = (S_n - np)\big/\sqrt{n \cdot p \cdot q}$. Cum este repartizată variabila S_n^*? Determinaţi experimental: $P[\, -k < S^*_n < k\,]$, unde k=1, 2, 3, 4, iar n şi p sunt constante.
Exerciţiu 6: Să se determine probabilitatea ca o variabilă aleatoare repartizată binomial, cu caracteristicile p=0.2, n=100, să se afle în intervalul[2,10]. Se vor considera eşantioane statistice de dimensiuni n = 10; 100; 250; 500.