Curs 5: Distribuţii clasice de probabilitate: Bernoulli, binomială, geometrică şi Poisson.

1.1 Distribuţia Bernoulli

Un experiment Bernoulli este un experiment aleator ce constă dintr-un număr precizat sau nu, de încercări independente. O încercare are doar două rezultate mutual exclusive, unul numit succes şi celălalt eşec. Variabila aleatoare, X, ce înregistrează rezultatul unei încercări se numeşte variabilă aleatoare Bernoulli. Exemplul clasic de experiment Bernoulli este aruncarea monedei: succesul fiind, de exemplu stema, iar eşecul, banul. Înregistrând 1 dacă se produce un succes şi 0 în caz de eşec, şi notând cu p probabilitatea succesului, distribuţia de probabilitate a variabilei aleatoare Bernoulli este:

$$X = \begin{pmatrix} 1 & 0 \\ p & 1-p \end{pmatrix} \tag{1}$$

Valoarea medie a unei variabile aleatoare Bernoulli este

$$M(X) = 1 \cdot p + 0(1-p) = p,$$

iar dispersia:

$$\sigma^2(X) = (1-m)^2 p + (0-m)^2(1-p) = (1-p)^2 p - p(1-p) = p(1-p)$$

.

Distribuţia Bernoulli se foloseşte în generarea de şiruri de biţi aleatori, sau la alegerea unei alternative din două.

Exemplul 1. (Run-uri de biţi) Având un şir de n biţi rezultaţi din simularea de n ori a variabilei aleatoare X de mai sus, numim run de biţi o succesiune de biţi identici în şir. De exemplu în şirul de biţi $0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0$, avem următoarele run-uri de biţi 1:

$$11, \quad 1, \quad 111$$

În analiza şirurilor de biţi aleatori folosiţi în probleme de securitate (criptografie) este foarte util numărul mediu de run-uri de biţi conţinuţi în şir.

Notăm cu N variabila aleatoare ce dă numărul de run-uri de biţi 1, în şirul de n biţi

$$b_1, b_2, \ldots, b_n,$$

obţinuţi din simularea (observarea) variabilelor aleatoare independente $X_1, X_2, \ldots, X_n$, Bernoulli distribuite, adică:

$$X_i = \begin{pmatrix} 1 & 0 \\ p & 1-p \end{pmatrix} \tag{2}$$

Practic simulăm aceeaşi variabilă aleatoare X, doar că îi asociem indicele $1, 2, 3, \ldots, n$ care indică al câtelea bit generăm.

Prin urmare fiecare bit b_i din şir este o "observaţie" asupra variabilei X_i, $i = \overline{1, n}$

Notăm cu Y_i variabila aleatoare ce ia valoarea 1 dacă în poziţia i a şirului de biţi începe un *run* de unu şi 0 în caz contrar. Astfel variabila aleatoare N, este suma variabilelor $Y_1, Y_2, \ldots Y_n$, $N = \sum_{i=1}^{n} Y_i$, iar numărul mediu de *run*-uri este $M(N) = \sum_{i=1}^{n} M(Y_i)$.

Dar conform definiţiei variabilelor Y_i avem că:

$$Y_1 = \begin{pmatrix} 1 & 0 \\ P(X_1 = 1) & P(X_1 = 0) \end{pmatrix}, \quad M(Y_1) = P(X_1 = 1) = p$$

adică un run de biţi începe din poziţia 1, dacă $X_1 = 1$. Pentru $i \geq 2$, un run de biţi începe din poziţia i, dacă bitul $b_i = 1$, dar bitul $b_{i-1} = 0$, adică pentru $i = \overline{2, n}$, $P(Y_i = 1) = P(X_{i-1} = 0, X_i = 1) = P(X_{i-1} = 0)P(X_i = 1) = (1-p)p$. Notând $\pi = (1-p)p$, variabila aleatoare Y_i are distribuţia de probabilitate:

$$Y_i = \begin{pmatrix} 1 & 0 \\ \pi & 1-\pi \end{pmatrix}$$

şi media $M(Y_i) = \pi = (1-p)p$. Prin urmare numărul mediu de *run*-uri de 1, în şirul de biţi $b_1, b_2, \ldots, b_n$, este $M(N) = p + (n-1)(1-p)p$.

Analiza statistică a *run*-urilor dintr-un şir de biţi se efectuează pentru a evalua caracterul aleator al acestora. Unul din testele folosite în criptografie este testul *run*, care numără *run*-urile de 0 şi 1 ditr-un şir de biţi produs de un generator de biţi aleatori. Dacă numărul *run*-urilor de 1,2,3,4,5 biţi şi respectiv ≥ 6 biţi, nu intră în intervale prescrise, şirul este considerat nesigur.

1.2 Distribuţia binomială

O variabilă aleatoare binomială este asociată unui experiment Bernoulli ce constă din n încercări (repetiţii). O încercare este o etapă a experimentului, ce are două rezultate mutual exclusive: unul numit *succes* şi celălalt *eşec*.

• Încercările sunt independente, în sensul că rezultatul unei încercări nu influenţează rezultatul celorlalte.

• Pentru fiecare încercare probabilitatea succesului este aceeaşi, p.

Exemple de experimente Bernoulli:

• Exemplul clasic este aruncarea de n ori a unei monede; succesul: cade banul, eşecul, cade stema;

• Observarea a n execuţii consecutive a grupului de instrucţiuni:

```
if (B) executã I1;
else executã I2;
```

Succesul= "expresia `B` este adevărată", iar eşecul= "expresia B este falsă".

• Transmiterea unei succesiuni de n biţi printr-un canal de comunicaţie. Succesul= "bit recepţionat corect", eşec="bit recepţionat incorect".

O variabilă aleatoare X asociată unui experiment Bernoulli, ce înregistrează numărul succeselor din n încercări se numeşte *variabilă aleatoare binomială*.

Mulţimea valorilor unei variabile aleatoare de tip binomial este $D = \{0, 1, 2, \ldots, n\}$, deoarece în n încercări se pot înregistra, 0 succese, un succes, ..., n succese.

Să determinăm distribuţia de probabilitate a unei variabile aleatoare binomiale X, în funcţie de numărul n al încercărilor şi probabilitatea p a unui succes într-o încercare, adică să calculăm probabilitatea ca în n încercări să înregistrăm k succese: $P(X = k)$, $k \in \{0, 1, 2, \ldots, n\}$.

În acest scop remarcăm că rezultatul fiecărei încercări este o observaţie asupra unei variabile aleatoare Bernoulli. Notăm cu $X_1, X_2, \ldots, X_n$ variabilele aleatoare Bernoulli, identic distribuite:

$$X_i = \begin{pmatrix} 1 & 0 \\ p & 1-p \end{pmatrix}, \quad i = \overline{1, n}$$

ce sunt "observate" în încercările $1, 2, \ldots, n$. Conform caracterizării experimentului Bernoulli, aceste variabile sunt independente şi deci vectorul aleator $(X_1, X_2, \ldots, X_n)$ are distribuţia de probabilitate dată de:

$$\begin{aligned} P((X_1, X_2, \ldots, X_n) &= (i_1, i_2, \ldots, i_n))) = P(X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n) \\ &= P(X_1 = i_1)P(X_2 = i_2) \cdots P(X_n = i_n) \\ &= p^k(1-p)^{n-k}, \end{aligned} \tag{3}$$

unde k este numărul de valori 1 în n–lista de biţi $(i_1, i_2, \ldots, i_n)$. Numărul de succese în n încercări este $X = X_1 + X_2 + \cdots X_n$. Evident că suma variabilelor Bernoulli $X_1 + X_2 + \cdots + X_n$ ia valori în mulţimea $\{0, 1, 2, \ldots n\}$. Probabilitatea ca în n încercări să avem k succese este

$$\begin{aligned} P(X_1 + X_2 + \cdots + X_n = k) &= P\Big(\bigcup_{\substack{(i_1,i_2,\ldots,i_n) \\ i_1+i_2+\cdots+i_n=k}} (X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n)\Big) \\ &= \sum_{\substack{(i_1,i_2,\ldots,i_n) \\ i_1+i_2+\cdots+i_n=k}} P(X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n), \end{aligned}$$

unde suma se calculează după toate n–listele de biţi $(i_1, i_2, \ldots, i_n)$ pentru care suma biţilor este k (există C_n^k astfel de liste!). Deci,

$$P(X_1 + X_2 + \cdots + X_n = k) = C_n^k p^k (1-p)^{n-k}$$

Ca să întelegeţi mai bine luăm un exemplu simplu $n = 3$ şi $k = 2$. În acest caz avem: $P(X_1 + X_2 + X_3 = 2) = P((X_1 = 1, X_2 = 1, X_3 = 0) \cup (X_1 = 1, X_2 = 0, X_3 = 1) \cup$
$\cup (X_1 = 0, X_2 = 1, X_3 = 1) = C_3^2 p^2 (1 - p)$

Conform formulei binomului lui Newton avem: $\sum_{k=0}^{n} C_n^k p^k (1-p)^{n-k} = (p+(1-p))^n = 1.$

În concluzie o variabilă aleatoare binomială este o variabilă discretă având distribuţia de probabilitate:

$$X = \begin{pmatrix} k \\ C_n^k p^k (1-p)^{n-k} \end{pmatrix}, \quad k = 0, 1, \ldots, n \tag{4}$$

Numele de variabilă aleatoare binomială vine deci de la faptul că $P(X = k)$ este termenul $k + 1$ al dezvoltării binomiale $(p + (1 - p))^n$.

Notăm cu $\text{Bin}(n, p)$ clasa variabilelor aleatoare binomiale, asociate unor experimente Bernoulli cu n încercări şi probabilitatea p a unui succes într-o încercare, iar $X \sim \text{Bin}(n, p)$ indică că X este o variabilă aleatoare ce are distribuţia de probabilitate binomială. n şi p se numesc parametrii distribuţiei binomiale.

Exemplul 2. După ce un virus a pătruns în sistem, se verifică starea a 20 de fişiere. Ştiind că probabilitatea ca un fişier să fi fost afectat de virus este 0.2, independent de celelalte fişiere, care este probabilitatea ca cel mult cinci din cele 20 de fişiere să fi fost deteriorate ?

Variabila aleatoare X ce dă numărul fişierelor infectate din cele 20, are distribuţia de probabilitate $\text{Bin}(n = 20, p = 0.2)$. Probabilitatea cerută este

$$P(X \leq 5) = \sum_{k=0}^{5} C_{20}^k (0.2)^k (0.8)^{20-k}$$

Exemplul 3. Presupunem că există 10 sateliţi GPS (Global Positioning System) pe orbită. O unitate GPS de la sol este activă, dacă cel puţin 4 sateliţi GPS funcţionează (pot fi contactaţi). Ştiind că încercările de contactare ale celor 10 sateliţi sunt independente şi că probabilitatea ca GPS-ul de la sol să eşueze în contactarea oricărui satelit din cei 10, este aceeaşi şi egală cu $p = 0.75$, să se calculeze probabilitatea ca unitatea GPS de la sol să fie activă.

Remarcăm că încercările de contactare a celor $n = 10$ sateliţi GPS definesc un experiment Bernoulli, cu $p = 0.75$ (în acest context succesul nu are sensul uzual, ci $p = 0.75$ este probabilitatea ca GPS-ul de la sol să eşueze în contactarea unui satelit).

Variabila aleatoare X, ce dă numărul de sateliţi ce nu pot fi contactaţi de la sol, are distribuţia de probabilitate $\text{Bin}(n = 10, p = 0.75)$. Astfel probabilitatea ca unitatea GPS de la sol să funcţioneze este probabilitatea evenimentului $(X \leq 6)$, adică probabilitatea ca cel mult 6 sateliţi să nu poată fi contactaţi (deci cel puţin patru să poată fi contactaţi):

$$P(X \leq 6) = \sum_{k=0}^{6} C_{10}^k (0.75)^k (0.25)^{10-k}$$

Propoziţia 1.2.1 *Fie X o variabilă aleatoare binomială $X \sim Bin(n, p)$. Media lui X este $M(X) = np$, iar dispersia, $\sigma^2(X) = np(1-p)$.*

Observaţia 1.2.1 *Experimentul Bernoulli se asimilează cu experimentul extragerii succesive a câte unei bile dintr-o urnă cu bile albe şi negre şi returnarea bilei după extragere. Asimilăm extragerea unei bile albe cu succesul şi a unei bile negre cu eşecul. Returnarea bilei după extragere, asigură independenţa încercărilor. Dacă bila nu se returnează, încercările nu mai sunt independente, deoarece conţinutul urnei se modifică după fiecare extragere.*

1.3 Distribuţia geometrică

Distribuţia geometrică se defineşte în contextul unei experiment Bernoulli în care numărul de încercări nu este fixat apriori. Presupunem că se efectuează o succesiune de încercări independente ce pot avea ca rezultat, succes sau eşec. Fie X_i rezultatul celei de-a i-a încercări. X_i ia valoarea 1, respectiv 0 după cum în a i-a încercare rezultatul este un succes, respectiv eşec. Notăm cu Y variabila aleatoare ce dă numărul de încercări până la primul succes, inclusiv.

- Mulţimea valorilor variabilei Y este $\mathbb{N}^* = \{1, 2, \ldots, n, \ldots\}$;
- Să determinăm distribuţia de probabilitate, adică să calculăm $P(Y = k)$, $k \in \mathbb{N}^*$: Evenimentul $(Y = k)$ se exprimă astfel

$$(Y = k) = (X_1 = 0, X_2 = 0, \ldots, X_{k-1} = 0, X_k = 1)$$

Dacă notăm cu p probabilitatea succesului, atunci datorită independenţei variabilelor $X_1, X_2, \ldots, X_k$, avem:

$$\begin{aligned} P(Y = k) &= P(X_1 = 0)P(X_2 = 0)\cdots P(X_{k-1} = 0)P(X_k = 1) \\ &= (1-p)^{k-1}p \end{aligned} \tag{5}$$

Să arătăm că suma seriei $\sum_{k=1}^{\infty}(1-p)^{k-1}p = 1$. Notând $q = 1 - p$, $m = k - 1$ avem seria geometrică cu raţia q ori p:

$$p\sum_{m=0}^{\infty} q^m = \frac{p}{1-q} = 1$$

Deci, o variabilă aleatoare Y ce are distribuţia geometrică, dă numărul de încercări într-un proces Bernoulli, până se obţine primul succes, inclusiv. Y ia o mulţime numărabilă de valori.

Notăm cu Geom(p) clasa variabilelor aleatoare ce au distribuţia geometrică de parametru p.

Dacă $X \sim$ Geom(p), X are distribuţia de probabilitate:

$$X = \begin{pmatrix} k \\ (1-p)^{k-1}p \end{pmatrix}, \quad k \in \mathbb{N}^*, p \in (0,1) \tag{6}$$

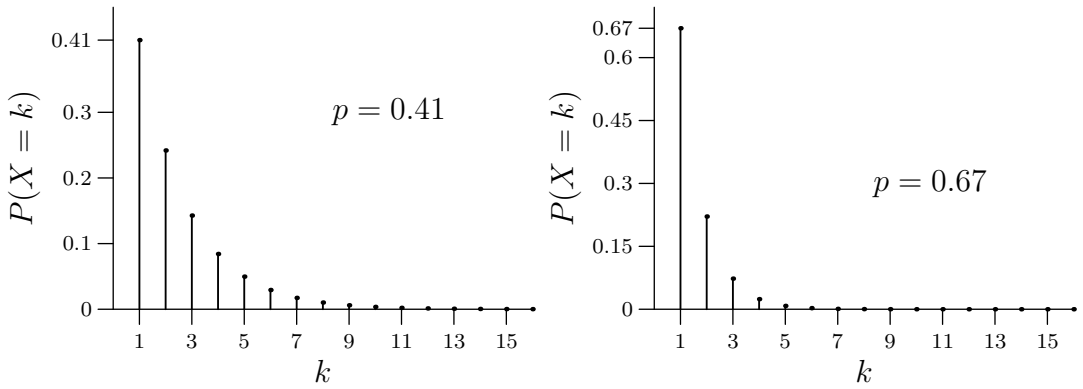

Fig.1: Ilustrarea distribuţiei de probabilitate geometrică pentru $p = 0.41$, respectiv $p = 0.67$.

Distribuţia de probabilitate geometrică, Geom($p = 0.41$), şi Geom($p = 0.67$), este ilustrată în Fig.1.

Propoziţia 1.3.1 *Dacă $X \sim Geom(p)$, atunci valoarea medie şi dispersia variabilei aleatoare, X, este:*

$$M(X) = \frac{1}{p}, \text{ iar } \sigma^2(X) = \frac{1-p}{p^2}. \tag{7}$$

Exemplul 4. Considerăm segmentul de program:

```
do{
bloc de instructiuni I;
  }while{!B}
```

Presupunem că expresia booleană `B` ia valoarea `true` cu probabilitatea p şi valoarea `false` cu probabilitatea $1-p$. Dacă rezultatele testelor succesive asupra lui `B` sunt independente ce distribuţie de probabilitate are numărul de execuţii ale grupului de instrucţiuni `I`?

Experimentul constă în parcurgerea a cel puţin o data a blocului I şi după fiecare parcurgere se înregistrează eşec dacă `B` este `false` (deci `!B` este `true`), respectiv succes în caz contrar. Variabila aleatoare N ce dă numărul de parcurgeri ale buclei `do-while` are distribuţia geometrică de parametru p. Într-adevăr, dacă de $k-1$ ori consecutiv, rezultatul testului este `!B este true`, iar după a k execuţie `!B este false`, atunci $P(N = k) = (1-p)^{k-1}p$.

Exemplul 5. Un algoritm generează la întâmplare un întreg fără semn, pe k biţi (k=32, de exemplu). Notaţia $\mathtt{s} \leftarrow_{\mathtt{R}} \{0,1\}^{\mathtt{k}}$ semnifică că algoritmul generează stringul de biţi `s`. La întâmplare înseamnă că fiecare string de k biţi poate fi generat cu aceeaşi probabilitate şi rezultatul fiecărei generări este independent de restul generărilor.

Dacă $s = (b_{k-1}b_{k-2}\ldots b_1b_0)_2$, este reprezentarea în baza 2 a unui întreg fără semn, atunci numărul zecimal asociat este $I(s) = \sum_{i=0}^{k-1} b_i\, 2^i$. Considerăm algoritmul probabilist:

```
do{
      s ←R {0,1}^k;
      n ← I(s);
}while (n ≥ M);
return n;
```

Să se determine distribuţia de probabilitate a variabilei aleatoare N, ce ia ca valori numerele pe k biţi returnate de algoritm şi distribuţia de probabilitate a variabilei, X, ce dă numărul de parcurgeri ale buclei `do-while`.

Rezolvare: Notăm cu Y variabila aleatoare ce are ca valori numerele întregi n ce pot fi generate de secvenţa:

$\mathtt{s} \leftarrow_R \{0,1\}^k$

$\mathtt{n} \leftarrow \mathtt{I(s)}$

Mulţimea valorilor lui Y în binar este $\{\underbrace{(00\ldots0)}_{k}, \ldots, \underbrace{(11\ldots1)}_{k}\}$, iar în baza 10:

$$D = \{0, 1, 2 \ldots 2^k - 1\}$$

Cardinalul lui D este $|D| = 2^k$. Deci probabilitatea de generare a unui număr întreg m, fără semn, pe k biţi, este $P(Y = m) = 1/2^k$, $\forall\ m \in D$. Variabila aleatoare N este variabila Y condiţionată de evenimentul $A = (Y < M)$, notat $N = (Y|Y < M)$.

Evenimentul $(Y < M) = (Y = 0) \cup (Y = 1) \cup \cdots \cup (Y = M - 1)$ are probabilitatea $P(Y < M) = P(Y = 0) + P(Y = 1) + \cdots + P(Y = M - 1) = M/2^k$. Variabila aleatoare N ia valorile $\{0, 1, 2 \ldots, M - 1\}$ şi

$$P(N = i) = P(Y = i|Y < M) = \frac{P(Y = i, Y < M)}{P(Y < M)} = \frac{1/2^k}{M/2^k} = 1/M,$$

$\forall\ i = 0, 1, \ldots M - 1$. Deci N are distribuţia uniformă pe multimea
$\{0, 1, 2, \ldots, M - 1\}$.

Variabila aleatoare, X, ce dă numărul de parcurgeri ale buclei este o variabilă aleatoare de tip geometric, cu probabilitatea de succes $p = P(Y < M) = M/2^k$.

1.4 Distribuţia Poisson

O variabilă aleatoare X, discretă, ce dă numărul de produceri ale unui eveniment rar, într-un interval de timp fixat se numeşte variabilă aleatoare de distribuţie Poisson.

Se presupune că evenimentele ce se pot produce în orice moment al intervalului de timp fixat sunt independente. Dacă evenimentul se produce cu "intensitate" constantă astfel încât, în medie, se produc λ evenimente în intervalul fixat de timp, atunci distribuţia de probabilitate a variabilei aleatoare Poisson este:

$$X = \begin{pmatrix} k \\ \dfrac{e^{-\lambda}\lambda^k}{k!} \end{pmatrix} \quad k = 0, 1, 2, \ldots n, \ldots, \tag{8}$$

adică probabilitatea ca evenimentul rar, A, să se producă de k ori într-un interval de timp fixat, este: $P(X=k)=\dfrac{e^{-\lambda}\lambda^k}{k!}$. Să verificăm că:

$$\sum_{k=0}^{\infty} P(X=k)=1. \tag{9}$$

Într-adevăr, $\displaystyle\sum_{k=0}^{\infty}\frac{e^{-\lambda}\lambda^k}{k!}=e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}$. Se ştie însă că $e^x=\displaystyle\sum_{k=0}^{\infty}\frac{x^k}{k!}$. Prin urmare,

$$\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}=e^{\lambda} \tag{10}$$

şi deci $\displaystyle\sum_{k=0}^{\infty}P(X=k)=1$.

Notăm cu Pois(λ) clasa variabilelor aleatoare Poisson de parametru λ.

Propoziţia 1.4.1 *Dacă X este o variabilă aleatoare de tip Poisson, cu parametrul λ, atunci media sa este $M(X)=\lambda$ şi dispersia $\sigma^2(X)=\lambda$.*

Variabilele aleatoare Poisson sunt singurele cunoscute, pentru care media este egală cu dispersia.

În Fig.2 se poate observa modul în care se schimbă distribuţia de probabilitate Poisson pe măsură ce se modifică parametrul λ.

Exemplul 6. Un server bază de date primeşte în medie 30 de cereri de acces pe secundă. Ştiind că cererile de acces sunt independente, ce distribuţie de probabilitate are variabila aleatoare, ce dă numărul acestor cereri/sec? Să se calculeze probabilitatea ca serverul să primească cel puţin o cerere în 6 milisecunde.

Observăm că variabila aleatoare X care dă numărul cererilor de acces pe secundă se asociază experimentului de observare a evenimentelor rare (deci care nu se produc în mod continuu) de accesare a bazei de date. Prin urmare variabila X are distribuţia Poisson de parametru $\lambda=30$ cereri/sec. O milisecundă este 1/1000 secunde. Dacă în medie există într-o secundă 30 de cereri, într-o milisecundă numărul mediu de cereri este $30/1000=0.03$, iar în 6 milisecunde de $\lambda'=6(0.03)=0.18$. Privind X ca o variabilă aleatoare ce are distribuţia Poiss(0.18), avem de calculat $P(X\geq 1)$. Dar cum X ia valori în $\mathbb{N}$, evenimentul $(X\geq 1)=\complement(X<1)=\complement(X=0)$ şi deci $P(X\geq 1)=1-P(X=0)=1-e^{-0.18}\dfrac{\lambda'^0}{0!}=1-e^{-0.18}$

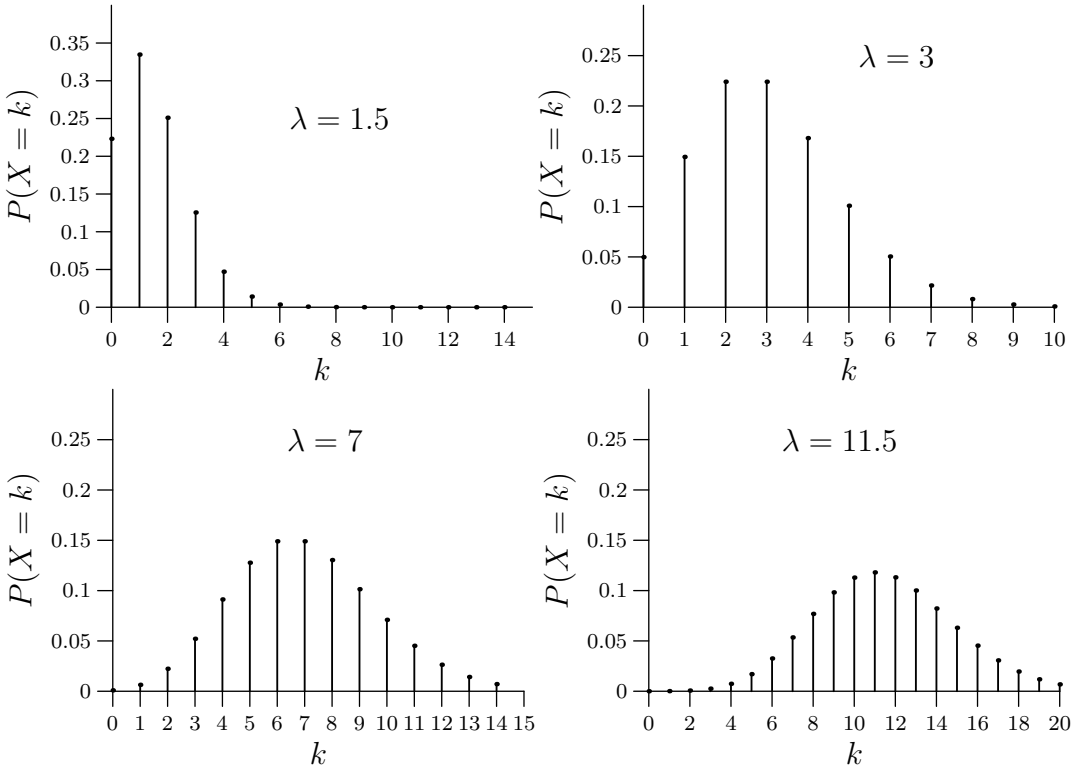

Fig.2: Vizualizarea distribuţiei Poisson pentru diferite valori ale parametrului λ.

Optional

1.5 Distribuţia Zipf

Distribuţia Zipf este o distribuţie discretă, ce a fost identificată în numeroase probleme de analiză a cererii de conţinut din WWW sau internet. De exemplu, distribuţia popularităţii paginilor WEB, distribuţia popularităţii fişierelor multimedia, descărcate de pe un server, sunt doar câteva exemple.

Pentru a întelege distribuţia Zipf luăm ca exemplu modalitatea de analiză a spamurilor, în scopul proiectării unui filtru spam. Se analizează mai multe emailuri spam şi se descompune textul acestora în unităţi, adică în termeni, cuvinte. Se determină apoi frecvenţa, f_i, a termenilor:

$$f_i = \frac{\text{numărul de câte ori apare termenul } t_i}{\text{numărul total de termeni identificaţi în toate spamurile analizate}}$$

Se ordonează descrescător lista termenilor, în raport cu frecvenţa, adică termenul $\mathtt{t_1}$ are frecvenţa maximă şi $\mathtt{t_n}$ are frecvenţa minimă. Indicele i în această ierarhizare se numeşte rangul termenului. S-a observat că distribuţia de probabilitate a variabilei aleatoare X ce asociază unui termen, rangul său, are distribuţia Zipf, adică probabilitatea ca rangul termenului să fie i este invers proporţională cu i:

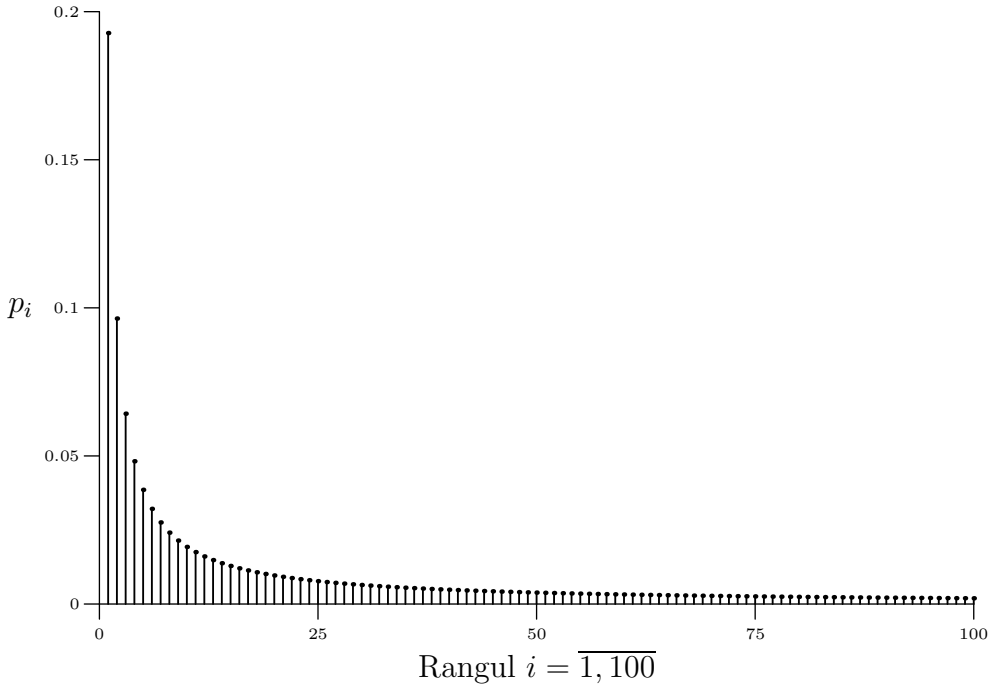

Fig.3: Distribuţia Zipf

$$P_X(i) = \frac{c}{i}, \quad i \in \{1, 2 \ldots, n\},$$

unde c o constantă de normalizare, ce se determină din condiţia:

$$\sum_{i=1}^{n} c/i = 1 \quad \Leftrightarrow \quad c = \frac{1}{\sum_{i=1}^{n} 1/i} = \frac{1}{H_n}$$

Deci variabila aleatoare X are distribuţia de probabilitate:

$$X = \begin{pmatrix} 1 & 2 & \ldots & i & \ldots & n \\ \dfrac{1}{H_n} & \dfrac{1}{2H_n} & \ldots & \dfrac{1}{iH_n} & \ldots & \dfrac{1}{nH_n} \end{pmatrix}$$

Numărul $H_n = \sum_{i=1}^{n} \dfrac{1}{i}$ se numeşte numărul armonic, de ordin n.

Astfel modelul probabilist al compoziţiei spam-urilor analizate este definit de variabila aleatoare:

$$X = \begin{pmatrix} t_1 & t_2 & \ldots & t_{n-1} & \ldots & t_n \\ \dfrac{1}{H_n} & \dfrac{1}{2H_n} & \ldots & \dfrac{1}{iH_n} & \ldots & \dfrac{1}{nH_n} \end{pmatrix}$$

unde $\mathtt{t_i}$, $i = \overline{1,n}$ sunt termenii cei mai frecvenţi din spam-uri, ordonaţi în ordinea descrescătoare a frecvenţei de apariţie.

În jurul anului 2000 s-au făcut primele investigaţii statistice relativ la WWW, în scopul de a deduce distribuţia de probabilitate a numărului de accesări ale paginilor WEB. Au fost monitorizaţi userii AOL din punctul de vedere al paginilor pe care le accesează într-o perioadă dată. S-a constatat că există puţine site-uri ce sunt vizitate de foarte mulţi surferi, în timp ce majoritatea sunt accesate foarte rar. Analizând datele rezultate din monitorizare s-a concluzionat că probabilitatea cererii de accesare a celei de-a i-a pagini, din ordinea descrescătoare a popularităţii, este invers proporţională cu cu i. Nivelul de popularitate a fost atribuit conform frecvenţei de accesare.

Interpretând reprezentarea grafică din Fig. 3 ca indicând probabilităţile de cerere de accesare a primelor 100 de site-uri monitorizate, se observă disproporţionalitatea exagerată în atragerea traficului, chiar la paginile 1-5.

Alte observaţii au condus la concluzia că cererea de accesare/descărcare de conţinut de pe internet are distribuţia Zipf generalizată, de parametru $\alpha \in (0, 5, 1]$. Numărul de accesări/descărcări ale fiecărui obiect depinde de popularitatea acestuia. Presupunem că obiectele sunt indexate descrescător în raport cu popularitatea, $O_1, O_2, \ldots, O_n$. Probabilitatea ca obiectul O_i să fie accesat/descărcat este $P(X = i) = \dfrac{C}{i^\alpha}$, adică cererea este caracterizată de o variabilă ce are distribuţia Zipf de parametru α:

$$X = \begin{pmatrix} 1 & 2 & \ldots & k & \ldots & n \\ \dfrac{C}{1^\alpha} & \dfrac{C}{2^\alpha} & \ldots & \dfrac{C}{k^\alpha} & \ldots & \dfrac{C}{n^\alpha} \end{pmatrix}, \quad C = \frac{1}{H_\alpha(n)},$$

unde

$$H_\alpha(n) = \frac{1}{1^\alpha} + \frac{1}{2^\alpha} + \cdots + \frac{1}{n^\alpha}$$

Pe baza acestor observaţii distribuţia Zipf generalizată este exploatată în proiectarea *cache*-ului (alegerea celor mai solicitate pagini WEB, documente, fişiere multimedia, pentru stocare în baza numită *cache*). Şi anume, dacă N este numărul mediu al accesărilor celor n obiecte de un anumit tip disponibile, într-un interval de timp dat, atunci numărul de accesări ale obiectului O_i este $N \cdot P(X = i) = \dfrac{N}{i^\alpha H_\alpha(n)}$. Numărul total de accesări a celor mai populare $k < n$ obiecte este:

$$\sum_{i=1}^{k} N \cdot P(X = i) = N \sum_{i=1}^{k} \frac{1}{i^\alpha H_\alpha(n)} = N \frac{H_\alpha(k)}{H_\alpha(n)}$$

Raportul dintre numărul de accesări ale primelor k, cele mai populare obiecte şi numărul de accesări ale tuturor celor n obiecte este:

$$h = \frac{N \dfrac{H_\alpha(k)}{H_\alpha(n)}}{N} = \frac{H_\alpha(k)}{H_\alpha(n)}$$

şi se numeşte *hit ratio*. Proiectarea bazei cache se face în aşa fel încât ea să aibă un raport h prescris. Din h şi α cunoscuţi se deduce $H_\alpha(k) = N \cdot H_\alpha(n)$, şi apoi k. k, astfel determinat, indică până la ce ordin de popularitate să fie inclus un obiect în *cache*.

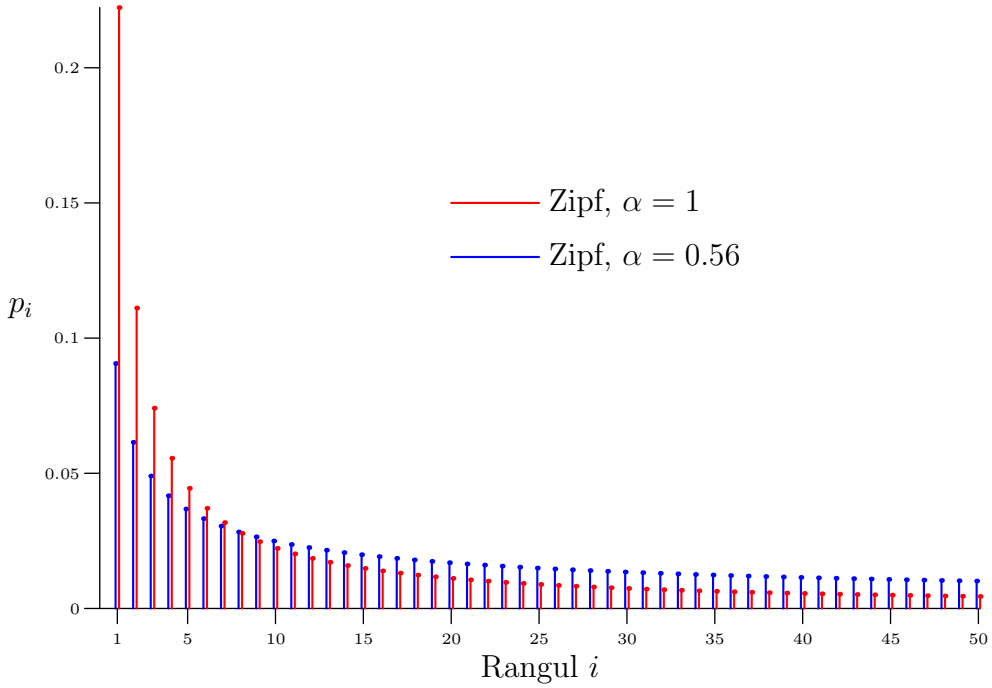

Fig.4: Distribuţie Zipf generalizată comparată cu cea clasică. Sunt trasate câte două segmente corespunzătoare aceleiaşi valori, unul pentru Zipf clasic (segmentul din dreapta) şi unul pentru Zipf generalizat (segmentul din stânga).

Pentru a putea proiecta motoare de căutare, de generaţie viitoare, este necesar să fie cunoscută în detaliu structura reţelei WWW. WWW este o reţea imensă, în continuă creştere. Aceasta creştere este aleatoare si neregulată: fiecare individ sau organizaţie putând sa-şi creeze site-uri WEB, cu un numar nelimitat de documente şi linkuri. WWW la un moment, t, este un graf G definit de nodurile, $V = \{1, 2, \ldots, n\}$, reprezentate de paginile WEB şi o multime de arce orientate $E \subset V \times V$, reprezentate de link-urile dintre pagini. Numărul de pagini (noduri în graf) se numeşte ordinul grafului, numarul de linkuri din pagina $i \in V$, se numeşte ordinul de ieşire din pagină, iar numărul de linkuri către pagina i, ordinul de intrare în acea pagină.

Topologia (structura) grafului WWW este dinamică, adica evoluează în timp, prin adaugarea şi/sau ştergerea unor pagini si link-uri. Astfel caracteristicile numerice, ca ordinul grafului şi gradele de intrare-ieşire din noduri (pagini) se modifică rapid. De aceea s-a ajuns la concluzia că cel mai potrivit model pentru graful WEB este cel de graf aleator.

Pentru a identifica caracteristicile unui astfel de graf, Barabási şi Albert au făcut observaţii asupra unui subgraf din WEB, şi anume asupra domeniul `nd.edu`, al Universităţii Notre Dame, Indiana, ce conţinea 325729 documente şi 1469680 linkuri. Experimentele au constat în crawling a subreţelei, la intervale aleatoare de timp. Principala caracteristică observată a fost că distribuţia de probabilitate a gradului de intrare, G_i,

respectiv iesire, G_o, într-o/dintr-o pagină, este una de tip Zipf generalizat, cu parametrii $\alpha = 2.1$, pentru distribuţia gradului de intrare, respectiv $\alpha = 2.45$, pentru G_o.

Pornind de la această observaţie experimentală s-a născut o nouă direcţie de studiu a grafurilor aleatoare, şi anume grafuri aleatoare de tip WEB. În Fig. 3 este ilustrată distribuţia Zipf clasică, iar în Fig.4 distribuţia Zipf generalizată, comparativ cu cea clasică.