

KITI PASISKIRSTYMAI, SUSIJĘ SU BERNULIO BANDYMAIS

1. Puasono proceso supratimas

Šį kartą kalba būtų apie vadinamuosius *retus* atsitiktinius įvykius – tokius, kurie per stebėjimui pasirinktą laiko tarpą (minutę, valandą, sekundę, dieną, savaitę, metus ir pan.) įvyksta **vidutiniškai** keletą (maždaug nuo 1 iki 10) kartų. Tie kartai ir yra stebimasis atsitiktinis dydis. Tikslas – nustatyti kiekvienos šio dydžio reikšmės (teoriškai tos reikšmės gali eiti nuo 0 iki begalybės) tikėtinumo mastą, tikimybę. Ši situacija ir jos teorinio modeliavimo būdas paprastai yra siejami su Deni **Puasono** (Poisson, 1781–1840) vardu. Pabrėžtina, jog svarbiausias parametras čia yra tų atsitiktinių įvykių *vidutinis intensyvumas*, t.y. jų pasirodymo per vienokį ar kitokį laiko tarpą (vienetą) *vidutinis kiekis* (vidurkis): šis vidutinis intensyvumas paprastai žymimas λ arba a . Jei tasai vidutinis intensyvumas (λ) išlieka stabilus, tai tokį palyginti retų atsitiktinių reiškinų srautą įprastai yra vadinti tiesiog *Puasono procesu*. Neblogi Puasono proceso pavyzdžiai galėtų būti, sakysim, gaisrinės išskvietimų skaičius per pamainą (8 val.), eismo avarijų įvykusių kokiame nors regione per konkretų laiko tarpą, skaičius, naujagimių, gimusių kokioje nors apylinkėje per tam tikrą laiką (savaitę, mėnesį ir pan.), skaičius ir t.t.: visų šių atsitiktinių dydžių pasiskirstymui modeliuoti labiausiai tiktų D. Puasono pasiūlytas „mechanizmas“. O jo esmė tokia: jeigu yra žinomas kokio nors reto atsitiktinio įvykio vidutinis intensyvumas (kitaip – pasirodymų per pasirinktą laiko tarpą skaičius) λ , tai tikimybė, jog per tą laiką šis įvykis įvyks m kartų, yra:

$$P_{(m; \lambda)} = \frac{\lambda^m}{m!} e^{-\lambda}$$

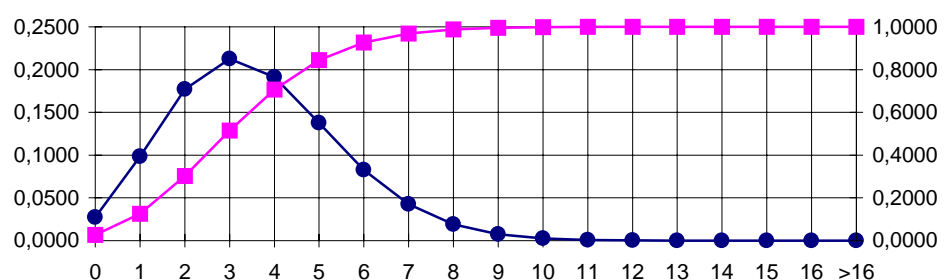
Čia m – „rūpinimas“ tiriamojo įvykio pasirodymų (per pasirinktą laiką) skaičius; $e \approx 2,71828...$ (Neperio skaičius), o λ , kurį būtina žinoti iš anksto ir kuris kaip tik ir lemia kiekvieną m atitinkančią tikimybę, yra išankstinis šio pasiskirstymo parametras, lygus vidutiniam tiriamųjų įvykių intensyvumui.

Pasiskirstymas (lentelė), vaizduojantis galimas m reikšmes, nurodytu būdu apskaičiuotas jas atitinkančias tikimybes $P_{(m)}$ bei pasiskirstymo funkcijos reikšmes $F_{(m)}$ yra vadinamas *Puasono pasiskirstymu*.

Pavyzdys. Siuvant mašina, siūlas nutrūksta vidutiniškai 3,6 karto per valandą. Reikia sudaryti siūlo nutrūkimų skaičiaus per val. (jį žymėsime įprastai : m) pasiskirstymą. Turėtume:

$m =$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	>16
$P_{(m)}$	0,0273	0,0984	0,1771	0,2125	0,1912	0,1377	0,0826	0,0425	0,0191	0,0076	0,0028	0,0009	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000
$F_{(m)}$	0,0273	0,1257	0,3027	0,5152	0,7064	0,8441	0,9267	0,9692	0,9883	0,9960	0,9987	0,9996	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000

Grafinis to pasiskirstymo vaizdas pateiktas 1 pav. (čia kreivės $P_{(m)}$ lūžio taškai paženklinėti rutuliukais, $F_{(m)}$ – kvadratai; $P_{(m)}$ reikšmės vaizduoja kairioji Y-ų ašis, $F_{(m)}$ – dešinioji):



1 pav. Puasono pasiskirstymas ($\lambda = 3,6$)

Kalbininkui Puasono procesu (pasiskirstymu) *tiesiogiai* irgi labiausiai tiktų modeliuoti kaip tik kalbėsenoje, kalbos sraute pasitaikančias palyginti retas atsitiktinines: tarties ar kirčiavimo klaidų, netinkamos žodžių vartosenos, skaitymo riktų, kalbėsenos „užsikirtimų“ ir pan. dalykų kiekius per pasirinktą laiko tarpą.

Pavyzdžiai:

- Skaitydamas paskaitą, lektorius per valandą vidutiniškai pavartoja 2,2 klausytojams nežinomos prasmės posakius. Reikia sudaryti jo vartojamų tokių posakių, pasakomų per val., kiekio pasiskirstymą.
- Diktorius, skaitydamas žinias, vidutiniškai kas 10 min. suklysta (susimaišo). Kokia yra tikimybė, kad per 30 min. trukmės žinių laidą jis suklys ne daugiau kaip du kartus?

Savaime suprantama, jog, susidūrus su Puasono procesu, atskiras klausimas yra tinkamo *laiko intervalo* (t.y. to laiko tarpsnio, per kurį ir fiksuojamas tyrinėtoją dominančių atsitiktinių įvykių skaičius) parinkimas. Šiaipjau, žiūrint tiek matematinės statistikos teorinių nuostatų, tiek ir skaičiavimo patogumo (formulėje figūruoja laipsninės funkcijos!), jį reiktų

parinkti taip, kad vidutinis jam „tenkančių“ tiriamųjų atsitiktinių įvykių kiekis (t.y. intensyvumas) nebūtų didelis (geriausia – kad λ neviršytų 10). Nors dirbant su šiuolaikiniais kompiuteriais bei programomis, skaičiavimo patogumas nebėra svarbiausias argumentas, šio reikalavimo vis viena dera laikytis.

Kartu pabrėžtina, kad būtina žiūrėti kito itin svarbaus momento – intensyvumo stabilumo, vienos iš pačių esmingųjų Puasono proceso sąlygų. Bet kita vertus – jo negalima ir absoliutinti: juk Puasono procesu „aprašomų“ atsitiktinių įvykių intensyvumas beveik niekuomet nebūna visą, neribotą laiką vienodas, pastovus amžinai, o irgi kinta priklausomai nuo konkrečių aplinkybių ir sąlygų (sakysim, vidutinis žmonių, per minutę įeinančių parduotuvėn, kiekis keičiasi priklausomai nuo paros meto). Galima būtų sakyti, kad esmingi šiaipjau vienintelio jų parametro – intensyvumo (λ) pokyčiai kaip tik ir *diferencijuoja* Puasono procesus, daro juos „skirtingus“.

2. Platesnis Puasono pasiskirstymo taikymas. Binominio pasiskirstymo apibendrinimas mažoms tikimybėms

„Grynas“ Puasono procesas paprastai nėra tiesiogiai siejamas su Bernulio bandymais. Bet reikia pridurti, kad jo esmę sudaranti tikimybių skaičiavimo formulė, pateikta anksčiau, gali būti ir kur kas plačiau taikoma: ne tik *laike* vykstančių, bet ir įvairiomis kitomis situacijomis pasireiškiančių (plg. gretimus pvz.) retų atsitiktinių įvykių tikėtinumui apskaičiuoti.

Pavyzdžiai:

Spaudai ruošiamą knygą. Viename jos lauzinio psl. vidutiniškai būna 2,74 korektūros klaidų. Kokia yra tikimybė, jog atsitiktinai paimtame šio lauzinio psl. korektūros klaidų bus daugiau kaip 3?

Tarkime, jog tyrimais nustatyta, kad vidutiniškai vienas žmogus iš 380-ties yra daltonikas. Miestelyje gyvena 4500 žmonių. Kokia yra tikimybė, kad dėl daltonizmo į vietos gydytoją kreipsis ne mažiau kaip 10 žmonių?

Be abejo, atsakyti į pvz. iškeltus klausimus reikėtų remiantis irgi Puasono pasiskirstymu: nors juose ir nekalbama apie *laike* vykstančius procesus, vis dėlto pagrindinis „pirminis“ duomuo bei, kartu, išankstinis pasiskirstymo parametras ir čia išlieka intensyvumas, vidutinis tiriamųjų atitiktinių reiškinių kiekis, tenkantis pasirinktam mato „vienetui“ (klaidų – puslapiui, susirgimų – tam tikram gyventojų skaičiui).

Grįžkime prie Bernulio bandymų. Teoretikai teigia, kad binominio pasiskirstymo formulė, pateikta ankstesnėje paskaitoje, nelabai tetinkanti ir tokiems Bernulio bandymų atvejams, kai tikimybės p (prisiminkime! p – rūpimo atsitiktinio įvykio tikimybė pavieniame bandyme!) reikšmė yra palyginti labai maža, o bandymų kiekis n – didelis. Tuomet vietoj binominio rekomenduojamas taikyti vėlgi *Puasono pasiskirstymas*. Tiktai tiriamojo atsitiktinio įvykio intensyvumas šiuo atveju nebūna iš anksto nurodomas „grynu pavidalu“, todėl yra apskaičiuojamas kaip tikimybės (pavieniame bandyme!) p ir atliktų bandymų skaičiaus n sandauga: $\lambda = np$. Neretai vietoj λ žymima $a = np$. Todėl tikimybės apskaičiavimo formulė dabar atrodytų taip:

$$P_{(m, np)} = \frac{(np)^m}{m!} \cdot e^{-np} \quad \text{arba, žymint } np = a, \quad P_{(m, a)} = \frac{a^m}{m!} \cdot e^{-a}$$

Natūralu, jog tokiais atvejais aktualiai kalbėti apie Puasono *procesą* kaip ir nebeiseitų: jis yra smarkiai modifikuotas, faktiškai – išvirtęs į Bernulio bandymų seriją.

Griežtos „ribos“, nurodančios, kuomet atsitiktiniams dydžiams, išplaukiantiems iš Bernulio bandymų, derėtų taikyti „klasikinį“ binominį ir kuomet – Puasono pasiskirstymą, kaip ir nėra. Bet kartais matematinės statistikos knygose pabrėžiama, jog Puasono pasiskirstymas tinkas tuomet, kai sandauga np esanti mažesnė už 10 (kitur gi ta riba nurodoma esanti lygi 4). Galima turbūt manyti, jog kuo sandauga np mažesnė, tuo labiau tinkas ir Puasono pasiskirstymas.

Lingvistinis pavyzdys:

Puasono pasiskirstymu galima būtų modeliuoti labai ilgų, *daugiau kaip 6 skiemenis turinčių* žodžių kiekio pasiskirstymą nemažos apimties tekstuose. Tokių žodžių santykinis dažnumas (ir – esama patikimo pamato spėti – tikimybė) rišliame tekste, remiantis preliminariais apskaičiavimais, yra **0,0011** (plg. *Kalbotyra*, t. 41(1), p. 38).

Tarkime, jog turime tekstą, kuriame iš viso yra 2634 žodžiai. Rūpi apskaičiuoti, kokios yra teorinės tikimybės, kad labai ilgų (ilgesnių nei šešiaskiemeniai) žodžių, jame nebus nė vieno, bus vienas, du, trys, keturi ir t.t.

Įsidėmėtina, kad Puasono pasiskirstymo atveju apie n ir p kaip *skirtingus* dydžius ar skirtingus parametrus aktualiai nebekalbama: parametru čia tampa *vienas* dydis a (arba λ), lygus jų *sandaugai*.

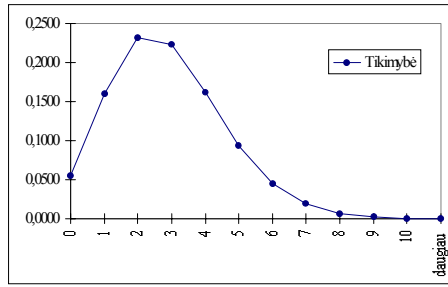
Grįžkime prie pavyzdžio apie labai ilgų žodžių kiekį tekste iš 2634 žodžių. Čia parametras

$$\lambda = a = np = 2634 \cdot 0,0011 = 2,8974.$$

Tada Pagal Puasono pasiskirstymo formulę apskaičiuotos tikimybės, atitinkančios labai ilgų žodžių kiekius, būtų tokios, kaip pavaizduota šioje pasiskirstymo lentelėje:

Kiekis (m):	0	1	2	3	4	5	6	7	8	9	10	daugiau
$P_{(m)}$:	0,0552	0,1598	0,2316	0,2236	0,1620	0,0939	0,0453	0,0188	0,0068	0,0022	0,0006	0,0002
$F_{(m)}$:	0,0552	0,2150	0,4466	0,6702	0,8322	0,9261	0,9714	0,9902	0,9970	0,9991	0,9998	1,0000

O 2 pav. pavaizduotas to paties pasiskirstymo grafikas (poligonas): x-sų ašyje atidėtas labai ilgų žodžių kiekis tekste iš 2634 žodžių, o y-ų ašyje – tų kiekių tikimybės. Kaip matome, ta tikimybė iš pradžių gana sparčiai auga, o vėliau – vėlgi staigiai mažėja, ir teorinė tikimybė, kad ilgų žodžių tokiam tekste bus daugiau negu 10 (vadinasi, nuo 11 iki pat 2634), tampa labai maža, tik 0,0002. Tai natūralu: kadangi labai ilgi žodžiai yra ir labai reti (tikimybė, kad atsitiktinai paimtas bet kuris teksto žodis bu labai ilgas, tėra tik 0,0011), tai daug jų rasti tekste – nėra ko tikėtis.



2 pav. Labai ilgų žodžių pasiskirstymas tekstuose iš 2634 žodžių

SVARBU:

Pagal Puasono dėsnį pasiskirsčiusio atsitiktinio dydžio ir *vidurkis*, ir *dispersija* taip pat yra lygūs sandaugai np , vadinasi, tokie pat, kaip ir parametras λ bei, tuo pačiu, lygūs tarpusavyje. Tai – svarbi šio pasiskirstymo ypatybė. Todėl jeigu eksperimentų metu gautų empirinių (imties) duomenų, kurių santykinis dažnumas nedidelis (praktiškai – mažesnis už 0,1), vidurkis ir dispersija yra lygūs tarpusavyje bei sandaugai np (arba tesiskiria nežymiai), tai teoriniam tokio pasiskirstymo modeliavimui greičiausiai tiktų kaip tik Puasono pasiskirstymas.

Tikimybės $P_{(m)}$ bei pasiskirstymo funkcijos $F_{(m)}$ reikšmėms apskaičiuoti Puasono pasiskirstymo atveju *Excel*’yje yra numatyta speciali funkcija

POISSON(m; lambda; kumuliatyvumas)

Jos sintaksė ir vartoseną visiškai analogiška „ankstesnės“ funkcijos *BINOMDIST* ($m; n; p; kumuliatyvumas$) sintaksei ir vartosenai: tesiskiria tik parametrai. Taip pat visiškai analogiškos yra ir pasiskirstymo funkcijos $F_{(m)}$ panaudojimo galimybės tuomet, kai reikia nustatyti tikimybės, kad m neviršys kokio nors „slenksčio“ k [$P_{(m \leq k)} = F_{(k)}$], kad jį viršys [$P_{(m > k)} = 1 - F_{(k)}$] arba bus ne mažesnis už tą slenkstį ($P_{(m \geq k)} = 1 - F_{(k-1)}$); o taip pat – kad m reikšmė papuls į intervalą nuo a iki b imtinai [$P_{(a \leq m \leq b)} = F_{(b)} - F_{(a-1)}$].

3. Puasono pasiskirstymo modifikacija: Čebanovo–Fukso pasiskirstymas

Filologams įdomi turėtų būti palyginti mažai žinoma Puasono pasiskirstymo modifikacija, vadinama Čebanovo–Fukso pasiskirstymu (žr. Piotrovskij R. G. ir kt. *Matematičeskaya lingvistika*, Moskva, 1977, p. 190–193). O įdomi ji tuo, kad čia galimų m reikšmių skalė yra tarsi paslinkta per vieną poziciją į dešinę; kitais žodžiais tariant, šiuo pasiskirstymu yra modeliuojamos situacijos, kuomet galimos m reikšmės prasideda ne nuo nulio, bet nuo vieneto. Filologams tai aktualu, nes didžiama lingvistinių objektų paprastai būna sudaryti mažiausiai bent jau iš vieno kokio nors elemento – sakysim, žodis būtinai turi bent jau vieną skiemenį, sakinys negali būti mažesnis, kaip vienas žodis, skiemuo taip pat turi turėti bent vieną garsą (ar – raidę) ir t.t. Tad Čebanovo–Fukso pasiskirstymas kaip tik ir leistų konstruoti tokio pobūdžio atsitiktinių dydžių pasiskirstymo teorinius modelius.

Analitinė Čebanovo–Fukso pasiskirstymo išraiška (kitais žodžiais – formulė) panaši į Puasono pasiskirstymo išraišką ir konkrečiai atrodytų taip:

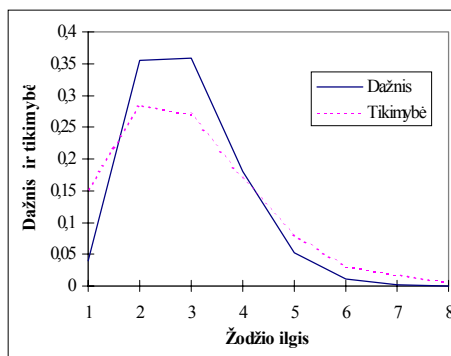
$$P_{(m+1)} = \frac{(a-1)^{m-1}}{(m-1)!} \cdot e^{-(a-1)}$$

Lingvistinis pavyzdys, pailiustruojantis Čebanovo–Fukso pasiskirstymo pritaikymą, galėtų būti vėlgi žodžių ilgio (skiemėnėmis) pasiskirstymas. Iš *Kalbotyros* 41(1) sužinome (žr. p. 38), kad rišliuose lietuviškuose tekstuose žodžiai pagal skiemenų kiekį pasiskirstę taip, kaip pavaizduota tolesnės lentelės 1–3 stulpeliuose (žodžių ilgio empirinis pasiskirstymas):

m	$P^*_{(m)}$	$F^*_{(m)}$	$P_{(m)}$	$F_{(m)}$
1	2	3	4	5
1	0,0401	0,0401	0,1513	0,1513
2	0,3553	0,3954	0,2858	0,4371
3	0,3594	0,7548	0,2698	0,7069
4	0,1802	0,9350	0,1698	0,8767
5	0,0527	0,9877	0,0802	0,9569
6	0,0111	0,9988	0,0303	0,9871
7	0,0010	0,9999	0,0095	0,9967
8	0,0001	1,0000	0,0026	0,9992

Taip pat yra žinomas (galimas apskaičiuoti) ir *vidutinis* žodžio ilgis (kuris atitiktų λ arba a): jis lygus 2,88822. Tad remiantis juo galima pagal pateiktą Čebanovo–Fukso pasiskirstymo formulę apskaičiuoti *teorinį* žodžio ilgių pasiskirstymą (t.y. $P_{(m)}$ bei $F_{(m)}$). Kitaip tariant, galima gauti teorinį (hipotetinį) to pasiskirstymo modelį, sukonstruotą remiantis prielaida, jog

žodžių ilgis skiemėmis galbūt gali būti pasiskirstęs pagal Čebanovo–Fukso dėsnį. Prielaidos pagrįstumą (tikėtumą) įvertintume specialiais būdais (kriterijais) lygindami teorines tikimybes su atitinkamais santykiniais dažnumais.



3 pav. Žodžių ilgio empirinis (dažnis) ir teorinis (tikimybė) pasiskirstymai

Grafiškai tos pačios lentelės informacija pavaizduota greta pateiktu paveikslėliu. Jame ištisine linija nubrėžtas santykinio dažnumo, o punktyrine linija – pagal Čebanovo–Fukso pasiskirstymo formulę apskaičiuotos teorinės tikimybės poligonas (grafikas). Ar tas net „plika akimi“ matomas tų dviejų grafikų skirtumas liudija ką tik suformuluotos prielaidos naudai, ar nenaudai – kitas klausimas, kurio šį kartą giliau dar nesvarstysime.

4. Geometrinis pasiskirstymas

Šis diskretusis pasiskirstymas priklauso vadinamajai *neigiamų* binominių pasiskirstymų grupei (šeimai), kurie visi yra susiję su Bernulio bandymais, tačiau, lyginant su binominiu pasiskirstymu, bendroji tų bandymų schema čia būna tartum „apversta aukštyn kojom“, atvirkščia: Bernulio bandymai kartojami tol, kol norimą kartų kiekį k bus gautas tiriamasis atsitiktinis įvykis, o bendras atliktų bandymų skaičius m ir yra jais modeliuojamas atsitiktinis dydis.

Geometrinis vadinamas bendro Bernulio bandymų kiekio (m) pasiskirstymas, kai tie bandymai kartojami tol, kol rūpimas atsitiktinis įvykis (pavieniame bandyme turintis tikimybę p) įvyksta vieną (pirmą) kartą.

Geometriniam pasiskirstymui būdingos šios ypatybės:

- * Išlaikomos pagrindinės Bernulio bandymų sąlygos, t.y.
 - įvykių baigmės - binariškos
 - įvykių baigmių tikimybės (p – teigiamos baigmės, $q=(1-p)$ – neigiamos) visą laiką išlieka pastovios
 - bandymai nepriklausomi; vieno jų rezultatas niekaip neįtakoja kito rezultato
- * bandymai kartojami tol, kol gaunamas (pirmą kartą) teigiamas rezultatas; vadinasi $k=1$
- * atliktų bandymų kiekis m bus atsitiktinis dydis ir turės geometrinį pasiskirstymą su parametru p .
- * dydžio m teoriškai galimų reikšmių aibė **prasideda nuo 1** ir eina iki „plius begalybės“

Dydžio m galimų reikšmių *tikimybės* yra:

$$P_{\{m, p\}} = p(1-p)^{m-1}$$

Dydžio m reikšmių, pasiskirsčiusių pagal geometrinį dėsnį, pagrindiniai parametrai yra tokie:

$$\text{teorinis vidurkis } m_{\text{vid}} = 1/p$$

$$\text{dispersija } D_m = (1-p)/p^2$$

Pavyzdžiai:

1. Šaudoma į taikinį; tikimybė pataikyti vienu šūviu (p) yra, tarkim, 0.64. Šaudoma tol, kol taikiny bus kliudytas. Iššautų šūvių kiekis ir bus geometrinį pasiskirstymą turintis atsitiktinis (m), galintis įgauti bet kokią reikšmę, pradedant 1 (kai taikiny kliudomas pirmuoju šūviu). Jo pasiskirstymo lentelė:

m :	1	2	3	4	5	6	7	8	9	10	11	12
$P(m, p)$:	0.64000	0.23040	0.08294	0.02986	0.01075	0.00387	0.00139	0.00050	0.00018	0.00006	0.00002	0.00001
$F(x)$:	0.64000	0.87040	0.95334	0.98320	0.99395	0.99782	0.99922	0.99972	0.99990	0.99996	0.99999	1.00000

Taikiniui kliudyti reikalingų šūvių *vidutinis kiekis* (vidurkis) bus $1/0.64=1.5625$; dispersija $D_m=(1-0.64)/0.64^2=0.8789$.

2. Urnoje yra 15 baltų rutuliukų ir 10 juodų. Rutuliukai su gražinimu traukiami tol, kol papuls juodas; bendras traukimų kiekis (skaičius) tad ir yra geometrinį pasiskirstymą (su parametru $p=10/25=0.4$) turintis atsitiktinis dydis (m), kurio teoriškai galimos reikšmės yra 1, 2, 3, Jo pasiskirstymo lentelė būtų tokia:

m :	1	2	3	4	5	6	7	8	9	10	11	12	>12

$P(m, p):$	0.40000	0.24000	0.14400	0.08640	0.05184	0.03110	0.01866	0.01120	0.00672	0.00403	0.00242	0.00145	0.00218
$F(x):$	0.40000	0.64000	0.78400	0.87040	0.92224	0.95334	0.97201	0.98320	0.98992	0.99395	0.99637	0.99782	1.00000

Vidutinis traukimų, reikalingų juodam rutuliukui ištraukti, skaičius (t.y. traukimų skaičiaus vidurkis) bus $1/0.4=2.5$; dispersija $D_m=(1-0.4)/0.4^2=3.75$.

Pastaba:

Gatavos funkcijos geometriškai pasiskirsčiusio dydžio reikšmių tikimybėms apskaičiuoti *Excel*is neturi: tą formulę tenka „užsiprogramuoti“ pačiam, pasitelkus laipsniams skaičiuoti skirtą funkciją POWER(pagrindas; rodiklis). „Konstrukcija“, apskaičiuojanti m_i reikšmės tikimybę $P(m_i)$, galėtų būti kad ir tokia:

$$= \text{POWER}((1-p_AbsAdr);(m_SantAdr-1))*p_AbsAdr$$

čia p_AbsAdr – absoliutinis ląstelės, kurioje laikoma p reikšmė, adresas
 $m_SantAdr$ – santykinis ląstelės, kurioje laikoma m_i reikšmė, adresas

Literatūra apie geometrinį pasiskirstymą:

- *Kruopis J.* Matematinė statistika, 1977, p. 47.
- *Ventcel E.S., Ovčarov L.A.* Teorija verovatnostej i ee inženernyje priloženija, 1988, p. 146-150.