ANALYSE EXPLORATOIRE DU DEGRÉ DE CONCORDANCE DES DONNÉES AVEC LA DISTRIBUTION BINOMIALE ET GÉOMÉTRIQUE SIMULÉE STATISTIQUEMENT EN PYTHON

Valentina ASTAFI

Département de Génie Logiciel et d'automatisation, groupe TI-231M, Faculté Ordinateurs, d'Informatique et de Microélectronique, Université Technique de Moldavie, Chisinau, République de Moldavie

Auteur correspondant: Astafi Valentina, valentina.astafi@isa.utm.md

Coordinateur scientifique: **Alexei LEAHU**, PhD, Prof., UTM

Résumé. *L'article présente les résultats d'une analyse exploratoire du degré de concordance des données de simulation statistique dans PYTHON par rapport aux distributions binomiale et géométrique, le volume des données simulées étant choisi en fonction de la précision et de la probabilité de confiance données pour la moyenne de sélection en tant qu'estimateur de la moyenne théorique. L'analyse est basée sur la comparaison graphique de la distribution de sélection des données avec la distribution théorique, mais aussi des caractéristiques numériques de sélection les plus représentatives avec les caractéristiques théoriques (moyenne, médiane, écart-type, coefficients d'asymétrie et d'aplatissement, etc.)*

Mots-clés: *analyse exploratoire, distribution binomiale, distribution géométrique, statistiques descriptives, simulations statistiques.*

Introduction

Les données jouent un rôle central dans la théorie des probabilités et les statistiques, car elles constituent la base sur laquelle on modélise et on comprend la complexité des phénomènes. Par exemple, les distributions binomiale et géométrique sont souvent utilisées pour modéliser des phénomènes discrets, des processus biologiques aux études de marché, mais aussi dans divers domaines scientifiques et d'ingénierie. Ces distributions fournissent des modèles mathématiques pour les phénomènes aléatoires qui sont fondamentaux pour comprendre les processus et les événements discrets. Cependant, l'application de la théorie dans la pratique nécessite une validation par la simulation et l'analyse exploratoire des données.

Même si ces distributions se ressemblent, elles modélisent des scénarios différents avec des caractéristiques et des utilisations différentes, allant de la modélisation des processus à la théorie de la décision. La distribution géométrique décrit le comportement probabiliste du nombre d'essais nécessaires pour obtenir le premier “succès”, tandis que la distribution binomiale indique le nombre de “succès” dans un nombre fixe d'essais indépendants. Cette caractéristique implique des différences significatives dans l'interprétation et l'utilisation de ces distributions dans la modélisation statistique et l'analyse des données.

Description des distributions

La distribution binomiale est une distribution de probabilité pour une variable aléatoire binomiale (a.v.), qui définit la probabilité d'obtenir un nombre fixe de succès dans un nombre fixe d'essais indépendants n, chacun avec la même probabilité p de succès. Cette distribution est souvent utilisée dans des situations où il n'y a que deux résultats possibles “succès” ou “échec” pour chaque essai, comme tirer à pile ou face, tester la qualité d'un produit, etc. Ces tests sont généralement appelés tests de Bernoulli [1].

La distribution binomiale est représentée par la formule (1), qui décrit la probabilité que le nombre total X de “succès” obtenus dans n essais de Bernoulli avec une seule et même probabilité de “succès” p dans chaque essai soit égal à k :

$$P(X = k) = (\frac{n!}{k!(n-k)!})p^k(1-p)^{n-k}\text{ , pour n'importe quel k=0,1 2,...,n.} \quad (1)$$

La distribution binomiale est applicable dans les situations où les événements sont indépendants, où la probabilité de "succès" est constante dans chaque essai et où l'intérêt réside dans le comptage des "succès"

La distribution géométrique (tronquée à zéro) décrit le nombre *total X* d'essais nécessaires pour obtenir le premier "succès" dans une série d'essais de Bernoulli avec une seule et même probabilité de "succès" *p* dans chaque essai [1]. C'est la seule distribution de type discret qui possède la propriété "sans mémoire" en ce sens que la probabilité que le premier "succès" soit enregistré dans l'échantillon avec *n+k,* sachant que n échantillons précédents étaient des "échec", ne dépend pas de *n*, mais seulement de *k*, et est calculée selon la même formule (2) pour la distribution géométrique :

$$P(X = k) = (1-p)^{k-1}p\text{ , pour n'importe quelle k=0,1,2,...} \quad (2)$$

Processus de simulation

Afin de comprendre et d'analyser le comportement probabiliste de la v.a. à travers le prisme du traitement de données simulées, on prendra 3 cas pour chaque distribution (Tab. 1), en faisant varier les paramètres *n* et *p* pour la distribution binomiale et p pour la distribution géométrique, où *n* est le nombre total d'essais, d'échantillons ou d'expériences, et *p* est la probabilité de "succès" dans chaque essai.

Tableau 1

Paramètres variables des distributions simulées

No.	Cas / Paramètres	Distribution binomiale		Distribution géométrique
		n	p	p
1	Cas 1	10	0,01	0,01
2	Cas 2	30	0,5	0,5
3	Cas 3	50	0,99	0,99

Après avoir choisi les valeurs des paramètres pour les trois cas de chaque distribution, calculez le nombre minimum de simulations nécessaires pour obtenir une estimation avec une certaine précision (ϵ=0,001) et une probabilité de confiance (par exemple 0. 95), en utilisant la formule dérivée du théorème central limite, qui est donnée par la formule (3), où z est la valeur du score z associée au niveau de confiance souhaité (par exemple, 1,96 pour 95 %), σ est l'écart type de l'a.v. X, et ϵ est la précision souhaitée (la distance maximale autorisée par rapport à la vraie moyenne de la population) :

$$n = (\frac{z \cdot \sigma}{\epsilon})^2 \quad (3)$$

En fait, ce nombre minimum de simulations coïncide avec la partie entière du nombre n augmenté d'une unité, puisque *n* peut être non entier. Pour appliquer cette formule, il faut estimer σ, l'écart-type de la population statistique de v.a. X, pour chaque distribution. Pour la distribution binomiale, l'écart-type (σ) est donné par la formule (4), où n est le nombre d'échantillons et p la probabilité de "succès".

$$\sigma = \sqrt{n \cdot p \cdot (1-p)} \quad (4)$$

L'écart-type d'une distribution géométrique est donné par la formule (5).

$$\sigma = \sqrt{\frac{1-p}{p^2}} \quad (5)$$

Par conséquent, on calcule le nombre minimum de simulations (tableau 2) pour chacun des trois cas de figure des deux distributions.

Tableau 2

Nombre de simulations

No.	Cas	Distribution binomiale	Distribution géométrique
1	Cas 1	380 318	38 031 840 000
2	Cas 2	28 812 000	7 683 200
3	Cas 3	1 901 592	39 196

Ces valeurs reflètent la nécessité d'ajuster le nombre de simulations en fonction des spécificités de chaque distribution et de ses paramètres afin d'atteindre le niveau souhaité de précision et de confiance dans l'analyse statistique. Le nombre significativement plus élevé de simulations requises pour la distribution géométrique avec p=0,01 souligne la sensibilité de ce type d'analyse aux paramètres de la distribution et à la précision souhaitée.

Afin de comprendre le comportement et les caractéristiques des distributions, il est nécessaire d'analyser non seulement la tendance centrale (identification de la ou des valeurs qui caractérisent le centre ou le point typique d'une distribution de données), mais aussi le coefficient d'asymétrie et de kurtosis.

Le coefficient d'asymétrie indique si la distribution est asymétrique d'un côté et à quel point [2]. Pour la distribution binomiale, le coefficient d'asymétrie ($\gamma 1$) est donné par la formule (6), qui donne une mesure de la symétrie de la distribution autour de sa moyenne.

$$\gamma 1 = \frac{1-2p}{\sqrt{n \cdot p(1-p)}} \quad (6)$$

Pour la distribution géométrique, le coefficient d'asymétrie est donné par la formule (7), illustrant l'asymétrie de la distribution des essais nécessaires pour obtenir le premier succès.

$$\gamma 1 = \frac{2-p}{\sqrt{1-p}} \quad (7)$$

Le coefficient de kurtosis indique le degré de concentration des données autour de la moyenne par rapport à une distribution normale [2]. Dans le cas d'une distribution binomiale, le coefficient ($\gamma 2$) est calculé selon la formule (8).

$$\gamma 2 = \frac{1-6p(1-p)}{n \cdot p(1-p)} \quad (8)$$

Pour la distribution géométrique, la valeur du coefficient est calculée selon la formule (9), indiquant la forme spécifique de la distribution des essais jusqu'au premier succès [2].

$$\gamma 2 = \frac{6+p^2}{1-p} \quad (9)$$

Le langage Python a été utilisé pour générer les simulations, à l'aide des bibliothèques NumPy, Matplotlib Pyplot et SciPy, qui fournissent des fonctions permettant de générer des nombres aléatoires selon des distributions binomiales et géométriques [3].

Analyse exploratoire et comparaison des distributions

Le traitement et la compréhension des ensembles de données sont rendus possibles par l'analyse exploratoire des données simulées, qui implique une série de techniques et de processus

d'exploration des données afin d'extraire des informations utiles. Pour chaque ensemble de données simulées, une analyse exploratoire est effectuée, en calculant : la moyenne, la médiane, la dispersion, l'asymétrie et le coefficient d'aplatissement pour chaque cas (Tab. 3 et Tab. 4) et en comparant ces valeurs avec les valeurs théoriques [4].

En analysant les données du tableau 3, on compare les moyennes et médianes théoriques avec les valeurs empiriques pour les distributions binomiale et géométrique, ce qui permet d'évaluer la précision et la fiabilité des simulations.

Tableau 3

Comparaison de la moyenne et de la médiane théoriques avec les valeurs simulées pour les distributions binomiale (1) et géométrique (2)

Nr.	Cas	Média		Médiane	
		théorique	***empirique***	***théorique***	***empirique***
1	Cas 1	0.1	0.0972	0.1	0.0
	Cas 2	15.0	15.0316	15.0	15.0
	Cas 3	49.5	49.4963	49.5	50.0
2	Cas 1	100	101.9626	69	70.5
	Cas 2	2.0	2.0247	1	2.0
	Cas 3	1.0101	1.0113	1	1.0

Dans le cas de la distribution binomiale, il n'y a que dans le premier cas qu'il y a une petite différence entre la moyenne et la médiane théoriques et la moyenne et la médiane empiriques, ce qui suggère une grande dispersion des données pour le petit nombre d'essais. Les cas 2 et 3 démontrent la précision des simulations entre toutes les valeurs théoriques et empiriques, avec un léger écart entre la médiane théorique et la médiane empirique.

Pour la distribution géométrique, le premier cas révèle une légère différence entre les moyennes théoriques et empiriques, mais une plus grande variance entre les médianes théoriques et empiriques, soulignant la variation naturelle du nombre d'essais requis pour un premier succès à faible probabilité. Dans les cas 2 et 3, des différences non significatives entre les valeurs théoriques et empiriques sont observées, ce qui indique l'efficacité de la modélisation de la distribution géométrique dans les scénarios avec des probabilités de succès modérées et élevées.

Le tableau 4 compare les valeurs théoriques de la dispersion, du coefficient d'asymétrie et du coefficient d'aplatissement avec celles obtenues à partir des simulations.

Tableau 4

Comparaison des valeurs théoriques avec les valeurs simulées: distributions binomiale (1) et géométrique (2)

Nr.	Cas	Dispersion		Coeff. d'asymétrie		Coeff. d'aplatissement	
		théorique	***empirique***	***théorique***	***empirique***	***théorique***	***empirique***
1	Cas 1	0.099	0.0969	3.1146	3.1871	9.5010	10.0013
	Cas 2	7.5	7.4658	0.0	-0.0389	-0.0667	-0.0448
	Cas 3	0.495	0.5074	-1.3929	-1.4305	1.9002	2.0804
2	Cas 1	9900.0	10722.8272	2.0000	2.1781	6.0607	7.6098
	Cas 2	2.0	2.0823	2.1213	2.1194	12.5	6.4327
	Cas 3	0.0102	0.0114	10.1	9.4933	698.0099	91.1076

En analysant les données de la distribution binomiale, on observe que le cas 1 présente des valeurs simulées proches des valeurs théoriques pour la dispersion et les coefficients d'asymétrie et d'aplatissement, bien qu'une asymétrie positive et un aplatissement plus élevé soient observés dans les données simulées. Le cas 2 est le plus proche des valeurs théoriques par rapport aux

valeurs simulées, ce qui indique la précision de la modélisation dans le cas d'une distribution équilibrée (p=0,5). Le cas 3 montre une légère différence entre la dispersion empirique et théorique, ainsi que de petites différences dans l'asymétrie et l'aplatissement, soulignant l'impact d'une forte probabilité de succès sur la forme de la distribution.

Pour la distribution géométrique, le cas 1 montre une faible différence entre la dispersion théorique et empirique, avec une asymétrie et un aplatissement plus élevé, reflétant la variabilité intrinsèque de la distribution lorsque la probabilité de succès est faible. Le cas 2 montre un alignement étroit entre les valeurs simulées et théoriques, indiquant la précision des simulations à une probabilité de succès modérée. Le cas 3 suggère que les extrêmes de la distribution géométrique peuvent être plus prononcés à des probabilités de succès très élevées (p=0,99). Les coefficients d'asymétrie et d'aplatissement augmentent de manière significative lorsque p s'approche de 1, ce qui indique une plus grande concentration des probabilités autour de valeurs faibles et la présence de rares valeurs aberrantes [5].

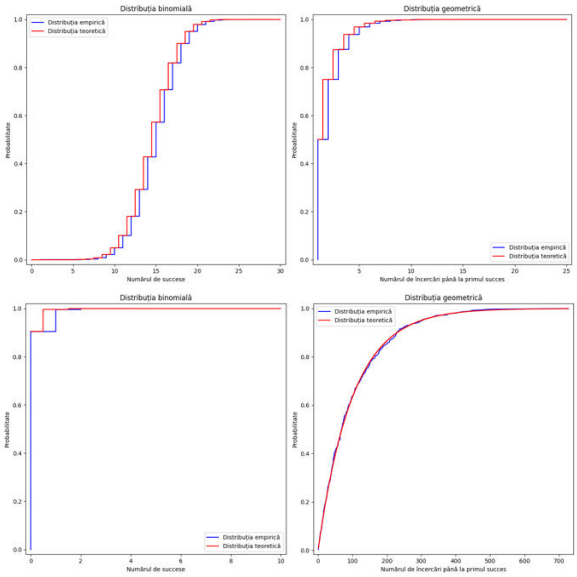
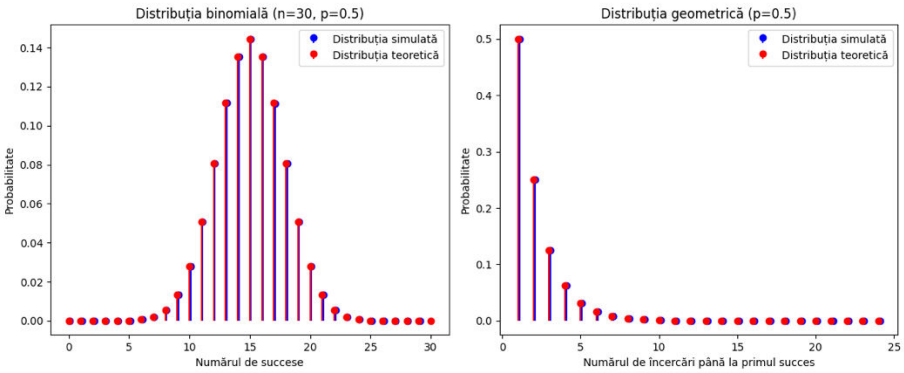


Figure 1. Graphique des fonctions de distribution binomiale et géométrique théoriques et simulées (1) cas n=30, p=0,5 et (2) n=10, p=0,01

La comparaison de la distribution cumulative théorique avec la distribution empirique basée sur un grand nombre d'essais permet de voir dans quelle mesure les données simulées correspondent aux prédictions théoriques. Ainsi, la figure 1 démontre la validité des modèles théoriques dans la description des phénomènes aléatoires, ainsi que la capacité d'utiliser des simulations pour approcher les caractéristiques de ces distributions.

Dans la figure 1 (2) n=10 et p=0,01 sur le graphique de la distribution binomiale, on observe des étapes correspondant à la fonction de distribution cumulative théorique et empirique. Les lignes se chevauchent presque, ce qui indique que les échantillons simulés correspondent bien à la distribution théorique attendue, étant donné que la distribution binomiale est bien définie pour un nombre fixe d'essais n et une probabilité de succès p pour chaque essai. Sur le graphique de la fonction de distribution géométrique, on voit que la distribution empirique suit la courbe de la distribution théorique. Le volume de données simulées est très important et le graphique tend déjà à s'aligner (les marches sont moins visibles). Comme p est très petit, on s'attend à ce que la plupart des valeurs soient concentrées vers le côté gauche du graphique, ce qui est également observé.

La Fig. 2 indique que les simulations reflètent les distributions théoriques correspondantes, ce qui démontre la fiabilité des simulations réalisées avec NumPy [3] pour modéliser des phénomènes aléatoires discrets.



=

Figure 2. Représentation graphique des distributions binomiale et géométrique théoriques et simulées (cas n=30, p=0,5)

Par rapport à la distribution binomiale, la distribution géométrique présente un comportement différent, mis en évidence par l'asymétrie et la forme du pic, en raison de la nature différente des processus qu'elle modélise. Cette analyse souligne l'importance de choisir le bon modèle de distribution en fonction de la nature des données et des phénomènes étudiés.

Conclusions

La comparaison des distributions binomiale et géométrique permettra non seulement d'explorer les propriétés spécifiques de chaque distribution, mais aussi d'apprécier l'importance de l'analyse exploratoire et des simulations dans la compréhension des phénomènes aléatoires. Ces observations soulignent la complexité de l'étude des probabilités et des statistiques, et leur applicabilité dans divers domaines. Les simulations permettent de comprendre le comportement des distributions et de vérifier la théorie mathématique par des expériences numériques. En outre, cet article présente une base pour une future comparaison de la qualité des générateurs pour différentes distributions de probabilité dans le contexte de leur mise en œuvre dans différentes applications (Python, R, Excel, Mathematica, Matlab, etc.) afin d'attirer l'attention sur l'amélioration de ces générateurs, éventuellement en proposant de meilleurs algorithmes de simulation statistique.

Remerciements

Je remercie sincèrement le PhD, Prof., Alexei Leahu pour le soutien et le temps accordés dans le processus de création de cet article. Grâce à ses conseils et sa vaste expérience dans le domaine des statistiques mathématiques : processus et applications aléatoires.

Références

[1] Geometric Distribution. In: *The Concise Encyclopedia of Statistics*. Springer, New York, NY., 2008, pp 226–227. doi: 10.1007/978-0-387-32833-1_164.

[2] Leahu A. Pârțachi I. *Probabilități, Statistică (Prin exemple și probleme propuse),* Partea I, Notițe de curs, Chișinău, Ed. ASEM, 2021, ISBN 978-9975-155-91-5.

[3] Applied Statistics in Python [En ligne]. Disponible : https://learn.saylor.org/course/view.php?id=504&sectionid=19783#section-25.

[4] Leahu A. *Analiza exploratorie a datelor*, Notițe de curs (format electronic), Chișinău, 2018.

[5] Grimmett G., Stirzaker D. *Probability and random processes.* Fourth Edition, Oxford university press, 2020. doi: 978-0-198-84759-5.