Séance 4
Les distributions statistiques

Honoka OYAMADA

Introduction

Cette séance porte sur les distributions statistiques et leur utilisation pour décrire, modéliser et comparer des phénomènes. Elle distingue distributions discrètes et continues, puis met en œuvre leur visualisation et le calcul de paramètres (moyenne, écart type) avec Python (scipy.stats, numpy, matplotlib). Des compléments théoriques (concentration, loi des grands nombres, théorème central limite) permettent de relier distributions et interprétation des résultats.

1. Questions de cours

1.1 Choisir entre variables discrètes et continues

Le critère principal est la nature de la variable :

Discrète : valeurs entières issues d'un comptage (nombre d'événements, occurrences). On représente une PMF (fonction de masse).

Continue : valeurs dans un intervalle (distance, durée, surface). On représente une PDF (densité), et les probabilités sont des aires sous la courbe.

Ce choix conditionne les outils (PMF vs PDF), l'interprétation des paramètres (moyenne/variance) et les méthodes de comparaison entre données observées et modèles.

1.2 Lois utiles en géographie

En géographie quantitative, certaines lois apparaissent régulièrement :

Normale : modèle de référence (symétrie, agrégation d'effets → TCL).

Log-normale / Gamma / Weibull : variables positives, asymétriques (tailles, durées, intensités).

Pareto / Zipf : distributions hiérarchisées et "queues lourdes" (tailles urbaines, inégalités territoriales).

Poisson : comptage d'événements rares.

χ^2, Student, Fisher : surtout pour l'inférence et les tests.

2. Mise en œuvre avec Python (ton code)
2.1 Ce que fait exactement le script

Liste de distributions (scipy.stats) + dossier de sortie images.

Fonctions :

mean_std() : calcule moyenne et écart type via dist.stats(moments="mv")

plot_discrete() : stem plot pour PMF

plot_continuous() : courbe pour PDF

Distributions tracées :

Discrètes : Dirac, uniforme discrète, binomiale, Poisson, Zipf(-Mandelbrot).

Continues : “Poisson continue” (approximation), normale, log-normale, uniforme continue, χ^2, Pareto.

2.2 Résultats numériques (moyenne et écart type)

(issus des paramètres exacts de ton code SciPy ; arrondis au millième)

Dirac en 0 : moyenne = 0 ; écart type = 0

Uniforme discrète {0,…,4} : moyenne = 2.000 ; écart type = 1.414

Binomiale (n=20, p=0,3) : moyenne = 6.000 ; écart type = 2.049

Poisson (λ=4) : moyenne = 4.000 ; écart type = 2.000

Zipf (a=2,0) : moyenne = ∞ ; écart type = ∞ (moments théoriques divergents à ce paramétrage)

Normale (μ=0, σ=1) : moyenne = 0.000 ; écart type = 1.000

Log-normale (s=0,5) : moyenne = 1.133 ; écart type = 0.604

Uniforme continue [0,1] : moyenne = 0.500 ; écart type = 0.289

χ^2 (k=4) : moyenne = 4.000 ; écart type = 2.828

Pareto (b=3) : moyenne = 1.500 ; écart type = 0.866

Remarque importante (rigueur) : ton code normalise la PMF Zipf sur un support fini (1..19) pour tracer ; en revanche stats() renvoie les moments théoriques de la loi Zipf, qui divergent pour $a \leq 2 \rightarrow$ d’où inf.

3. Analyse des documents + analyse des figures

3.1 Concentration : courbe de Lorenz et indice de Gini

Figure : Indice-de-Gini.png
Le graphique illustre une courbe de Lorenz et l’aire associée au calcul de l’indice de Gini : plus la courbe s’éloigne de la diagonale d’égalité parfaite, plus la distribution est concentrée (inégalitaire). Ce cadre est directement mobilisable en géographie (inégalités spatiales, concentration d’une population ou d’un revenu, hiérarchies territoriales).

3.2 Loi des grands nombres (LGN)

Doc : 02-Loi-des-grands-nombres.md
La LGN formalise la convergence de la moyenne empirique vers l'espérance lorsque la taille d'échantillon augmente. Méthodologiquement, cela justifie l'usage d'indicateurs (moyennes) et stabilise l'interprétation à mesure que le nombre d'observations croît.

3.3 Théorème central limite (TCL)

Doc : 03-Theoreme-central-limite.md
Le TCL explique pourquoi la loi normale intervient fréquemment : la somme (ou moyenne) de nombreuses variables indépendantes tend vers une distribution normale sous conditions générales. C'est un argument théorique de base pour utiliser la normale comme approximation.

4. Analyse des lois (à partir des courbes fournies)

Je commente chaque figure de façon courte, "rapport Forriez" (description → propriété → intérêt).

Lois "références"

Normale centrée réduite — Loi-normale-centree-reduite.png
Distribution symétrique autour de 0, maximum au centre, décroissance des deux côtés. Sert de référence (standardisation, TCL).

Normale : probabilité critique — Loi-normale-centree-reduite-Probabilite-critique-v2.png
Illustration d'une zone de rejet / région critique (aire en queue). Utilisée dans les tests (p-valeur, seuil α).

Lois à support borné

Uniforme — Loi-uniforme.png
Densité constante sur un intervalle : absence de préférence. Utile comme modèle simple ou hypothèse de base.

Triangulaire — Loi-triangulaire-Exemple.png
Support borné + mode unique : approximation simple quand on connaît min, max et valeur la plus probable.

Lois asymétriques positives (queues à droite)

Log-normale — Loi-log-normale-centree-reduite.png
Asymétrique, valeurs positives, queue à droite : adapté aux phénomènes multiplicatifs (tailles, distributions très étalées).

Gamma — Loi-gamma-Exemple.png
Positive, asymétrique ; souvent utilisée pour des durées/temps d'attente ou intensités.

Weibull — Loi-de-Weibull.png
Flexible (forme dépend des paramètres), utilisée pour durées/fiabilité, ou pour modéliser des variables positives non symétriques.

χ^2 (exemple) — Loi-du-chi-2-Exemple.png et Loi-du-chi-2-de-1-100.png
Positive, asymétrique ; dépend des degrés de liberté. Utilisée en tests (variance, ajustement). Ici, les deux figures illustrent surtout l'effet du paramètre (k faible → forte asymétrie ; k grand → forme plus "centrée").

Lois à queue lourde / hiérarchies

Pareto — Loi-de-Pareto-Exemple.png
Queue lourde : peu de valeurs très grandes existent mais pèsent beaucoup. Interprétation directe en géographie des hiérarchies (villes, revenus).

Cauchy — Loi-de-Cauchy.png
Queue très lourde : moyenne/variance non définies. Montre une limite importante : certains paramètres classiques (moyenne, écart type) peuvent devenir non pertinents.

Lois liées aux tests et extrêmes

Student — Loi-de-Student.png
Symétrique, plus "épaisse" que la normale en queue (selon ddl). Utilisée lorsque σ inconnue / petits échantillons.

Fisher (F) — Loi-de-Fisher-Exemple.png
Positive, asymétrique ; utilisée pour comparer des variances (tests F, ANOVA).

Gumbel — Loi-de-Gumbel.png
Loi d'extrêmes : modélise des maxima/minima (crues, températures extrêmes), utile pour risques et aléas.

Conclusion

La séance 4 met en place une typologie opérationnelle des distributions (discrètes/continues) et montre l'importance du choix de modèle pour interpréter des données géographiques. L'usage combiné des visualisations et des paramètres (moyenne, écart type) permet une première caractérisation, tout en rappelant que certains modèles (Zipf, Cauchy, Pareto selon paramètres) impliquent des limites sur les moments et donc sur l'interprétation statistique.