SÉANCE 2 : Les principes généraux de la statistique

I- Question de cours

1- Quel est le positionnement de la géographie par rapport aux statistiques ?

La géographie se tient souvent à distance des statistiques, méprisant ou oubliant son importance dans l'apport à la discipline, alors même que l'analyse statistique et de données est essentielle dans l'obtention de résultats, notamment du fait de la production de données massives par la discipline. Cela pousse à regrouper les statistiques sous un seul terme assez vaste : "l'information géographique".

2- Le hasard existe-t-il en géographie ?

Il n'existe pas de réponse complètement certaine à cette question, cela dépend beaucoup de la position (notamment philosophique) sur la question. Si l'on prend la position déterministe : le hasard n'existe pas car tout événement a une cause. Ce qu'on pourrait qualifier de hasard ne vient que de variables non identifiées. En revanche, la théorie du chaos admet le hasard comme cause cachée, mais il provient alors de notre ignorance. Plus précisément au regard de la géographie, le hasard statistique existe. En effet, on distingue deux types de hasard : le hasard bénin et le hasard sauvage, qui sont présents objectivement en géographie, ce qui pousse à n'admettre que des tendances. A cela s'ajoute le jeu de la multiscalarité : des tendances à petite échelle peuvent ne pas se retrouver à grande échelle. Finalement, on ne peut pas nier la présence du hasard en géographie notamment au regard des échelles, mais l'usage de la statistique reste indispensable à la discipline pour dessiner des tendances et structures et éviter au plus le hasard.

3- Quels sont les types d'information géographique ?

Il existe deux séries statistiques dans l'information géographique. En premier, les éléments de géographie humaine ou géographie physique (la base attributaire d'un système d'information géographique). En deuxième, la morphologie même des ensembles délimités (les données géométriques du S.I.G).

En parallèle de l'information géographique, on trouve aussi en amont la nomenclature qui correspond à un ensemble de définitions préalables au recueil de l'information, et les métadonnées qui permettent un examen critique des sources.

4- Quels sont les besoins de la géographie au niveau de l'analyse de données ?

La géographie a besoin de l'analyse de données pour confronter les données analysées et les résultats obtenus avec la méthodologie de production et avec ce que l'on connaît du phénomène étudié. L'analyse de données permet d'apporter des probabilités (pour déterminer les lois du hasard) et des statistiques (pour établir la bonne loi de probabilité) à la recherche. D'autant plus que la plupart du temps, le géographe ne produit pas ses données d'analyse.

5- Quelles sont les différences entre la statistique descriptive et la statistique explicative ?

La statistique descriptive consiste à étudier des données pour dégager des propriétés remarquables par rapport à une distribution théorique connue et donc obtenir une image simplifiée de la réalité. Cette statistique décrit les données de l'échantillon étudié. Tandis que

la statistique explicative se fonde sur les statistiques descriptives pour prédire. Finalement, la statistique descriptive se situe en amont de la statistique explicative.

6- Quelles sont les types de visualisation de données en géographie ? Comment choisir celles-ci ?

Les types de visualisations de données en géographie sont multiples : les visualisations graphiques comme un histogramme, une représentation sectorielle (diagrammes) ou encore des courbes, les divers tableaux d'analyse et de synthèse, mais en géographie on peut aussi penser à la cartographie.

On choisit les types de visualisations en fonction de plusieurs critères : si l'on possède des variables quantitatives ou qualitatives, mais aussi en fonction de la structure et de la distribution des données, du nombre de données observées et de ce que l'on souhaite démontrer ou expliquer.

7- Quelles sont les méthodes d'analyse de données possibles ?

Il existe trois méthodes d'analyse de données.

Tout d'abord les méthodes descriptives, qui permettent de visualiser et classer les données en résumant le tableau des variables. Cette méthode se fonde sur l'analyse factorielle en composantes principales, l'analyse factorielle des correspondances, l'analyse des proximités ou encore les méthodes classification.

Ensuite, les méthodes explicatives permettent de relier une variable à expliquer à des variables explicatives.

Enfin, les méthodes de prévision pour l'analyse et la prévision d'une série chronologique en reliant le présent au passé.

8- Comment définiriez-vous : (a) population statistique ? (b) individu statistique ? © caractères statistiques ? (d) modalités statistiques ? Quels sont les types de caractères ? Existe-t-il une hiérarchie entre eux ?

a- La population statistique est l'ensemble des données, par exemple le nombre d'habitants d'un territoire.

b- L'individu statistique est un élément de la population statistique, une unité statistique. En géographie on distingue deux particularités, ils sont localisables et cartographiables, et ils sont eux-mêmes composés d'éléments de niveau inférieur.

c- Les caractères statistiques sont les particularités d'individu pris dans la population statistique sur laquelle l'analyse porte. On distingue en fonction du nombre de caractères étudiés une série numérique à une dimension, à deux dimensions ou multidimensionnelle.

d- Les modalités statistiques sont les valeurs prises par un caractère pour définir l'appartenance ou la non appartenance d'un individu à une modalité.

Les types de caractères sont soit une variable qualitative, soit une variable quantitative, qui deviennent une valeur aléatoire pour la première ou une variable statistique pour la deuxième lors d'une étude statistique. On peut subdiviser une variable qualitative en deux avec les variables qualitatives nominales qui décrivent des états, et les variables qualitatives ordinales qui décrivent des relations. De même pour une variable quantitative on distingue les variables quantitatives discrètes pour compter les données, et les variables quantitatives continues pour mesurer.

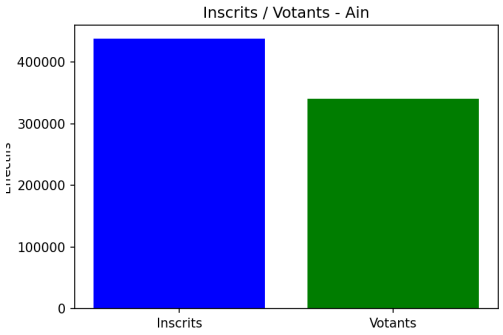
9- Comment mesurer une amplitude et une densité ?
L'amplitude se mesure en faisant la longueur b moins a. Avec a comme la valeur minimale de la classe et b la valeur maximale.
La densité se calcule en faisant le rapport (division) entre l'effectif *ni* et l'amplitude de la classe (*b* - *a*).

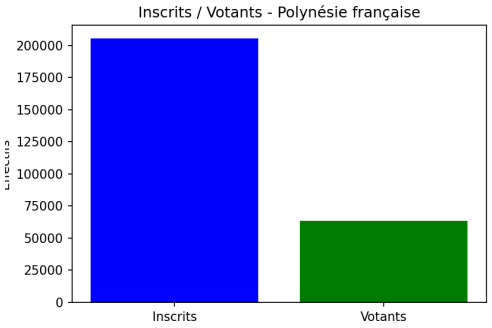
10- À quoi servent les formules de Sturges et de Yule ?
Les formules de Sturges et de Yule permettent de donner une valeur approximative du nombre de classes, car le nombre de classes ne doit être ni trop petit ni trop grand.

11- Comment définir un effectif ? Comment calculer une fréquence et une fréquence cumulée ? Qu'est-ce qu'une distribution statistique ?
- Un effectif correspond au nombre d'apparitions d'une variable dans la population.
- Pour calculer une fréquence, il faut faire le rapport entre l'effectif et l'effectif total. Et pour une fréquence cumulée, il faut faire la somme des fréquences associées aux valeurs inférieures ou égales à *k* (modalités), ou le rapport entre l'effectif cumulé et l'effectif total.
- Une distribution statistique associe les classes de valeur à leur fréquence d'apparition, cela permet de mettre en avant un type de loi de probabilité.

II- Mise en oeuvre avec Python

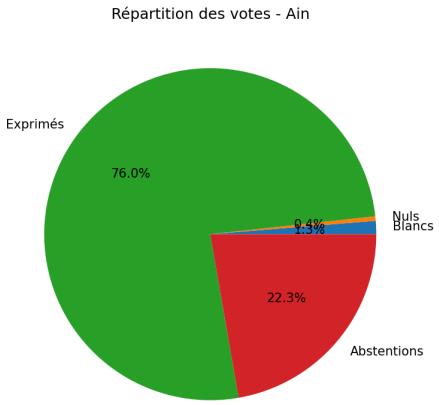
11- Diagrammes en barres avec le nombre des inscrits et des votants pour chaque département

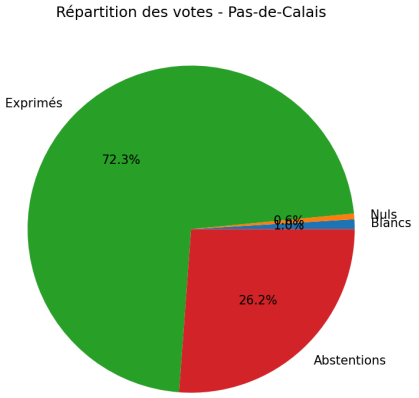


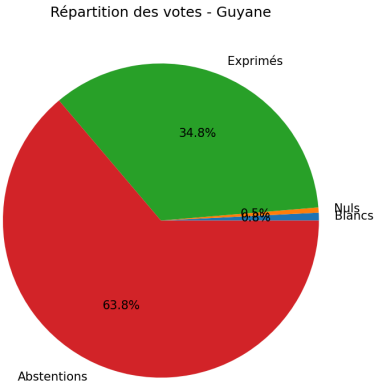


A partir de l'élaboration de ces diagrammes en barres, on peut mettre en avant la différence entre le nombre d'inscrits et le nombre de votants : les votants sont toujours inférieurs aux inscrits totaux dans un département. Mais ces diagrammes mettent aussi en lumière l'écart majeur avec les départements en Outre-Mer : le nombre d'inscrits reste généralement le même, mais le nombre de votants est beaucoup plus faible, la différence est donc d'autant plus marquée.

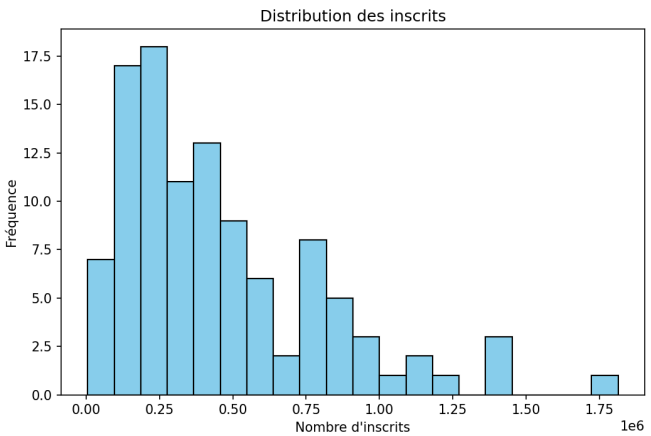
12- Diagrammes circulaires avec les votes blancs, nuls, exprimés et l’abstention pour chaque département







Ces diagrammes circulaires mettent cette fois-ci en lumière le rapport entre le nombre de votes exprimés et l'abstention dans chaque département. On peut tout d’abord voir que dans les départements en France hexagonale, l’abstention représente à peu près toujours un tiers, tandis que les votes exprimés deux tiers. Les votes nuls et blancs restent minimes. Mais à nouveau une grande différence s’opère avec les départements d’Outre-Mer, où l'abstention monte à deux tiers pour un tiers de votes exprimés.



SÉANCE 3 : Les paramètres statistiques élémentaires

I- Question de cours

1. Quel caractère est le plus général : le caractère quantitatif ou le caractère qualitatif? Justifier pourquoi.

Le caractère quantitatif est plus général : la grande majorité des paramètres statistiques concernent des variables quantitatives (comme les paramètres de positions, de dispersion ou de forme), tandis que cela ne concerne que ponctuellement les caractères qualitatifs, par exemple avec les fréquences.

2. Quels sont les caractères quantitatifs discrets et caractères quantitatifs continus ? Pourquoi les distinguer ?

Les caractères quantitatifs discrets sont des valeurs isolées et dénombrables, on peut donc faire le classement de ces valeurs par ordre croissant. Tandis que les caractères quantitatifs continus sont des valeurs appartenant à un intervalle. On les distingue car leurs formules ne sont pas les mêmes, par exemple pour une moyenne, une variable quantitative discrète est calculée par une somme, tandis qu'on utilise une intégrale pour une variable quantitative continue.

3. Paramètres de position

- **Pourquoi existe-t-il plusieurs types de moyenne ?**

Il existe plusieurs types de moyenne car cela varie en fonction de la nature de la variable : toutes les situations ne permettent pas une moyenne arithmétique. Chaque moyenne suppose différentes conditions (moyenne arithmétique, quadratique, harmonique, géométrique, mobile, fonctionnelle).

- **Pourquoi calculer une médiane ?**

Une médiane permet de diviser une série en deux parties exactement égales l'une à l'autre, ce qui lui vaut son nom de "moyenne du milieu". De plus, la médiane a l'atout de ne pas être influencée par les valeurs extrêmes à la différence de la moyenne arithmétique. Cela lui permet aussi de bien résumer des distributions fortement dissymétriques.

- **Quand est-il possible de calculer un mode ?**

Il est possible de calculer un mode lorsqu'une valeur est plus fréquente que d'autres, et qu'il existe donc aussi une valeur dominante ou un effectif maximal. On ne peut par exemple pas calculer un mode si toutes les valeurs ont la même fréquence.

4. Paramètres de concentration

- **Quel est l'intérêt de la médiale et de l'indice de C. Gini ?**

L'intérêt de la médiale est de partager en deux parties égales, représentant chacune 50% des valeurs globales, donc elle ne représente plus uniquement l'effectif comme la médiane mais aussi l'importance de la totalité du caractère possédé par les individus. L'intérêt de la médiale est aussi de la comparer avec la médiane pour avoir une mesure de concentration (forte/faible).

D'un autre côté, l'intérêt de l'indice de C.Gini est de pouvoir décrire les effets de la concentration d'une population statistique. En effet, cela se repère à partir d'une courbe et

de valeurs variant entre 0 et 1 : plus la courbe s'éloigne de la diagonale, plus la concentration est importante. Ainsi, l'indice de C.Gini permet une visualisation des inégalités d'une distribution.

5. Paramètres de dispersion

- **Pourquoi calculer une variance à la place de l'écart à la moyenne ? Pourquoi la remplacer par l'écart type ?**

Il vaut mieux calculer une variance car cette dernière tient compte de toutes les données, il s'agit donc de la meilleure caractéristique de dispersion, contrairement à l'écart à la moyenne. Mais il est plus pratique de la remplacer par l'écart type car il est exprimé dans la même unité que la moyenne : en effet la variance est exprimée dans une unité au carré, tandis que l'écart type est la racine carrée de la variance. L'écart type est donc plus facile à interpréter.

- **Pourquoi calculer l'étendue ?**

Calculer l'étendue permet d'obtenir l'écart entre la plus grande valeur observée et la plus petite, de plus cette dernière est facile à calculer.

- **À quoi sert-il de créer un quantile ? Quel(s) est (sont) le(s) quantile(s) le(s) plus utilisé(s) ?**

Créer un quantile permet de partager la série statistique en parties égales, à partir de ces derniers on peut évaluer la répartition interne des données. Les quantiles les plus utilisées sont les quartiles : ils permettent de partager la série en quatres parties égales et donc d'obtenir la médiane (Q2) ou encore l'écart et l'étendue interquartile.

- **Pourquoi construire une boîte de dispersion ? Comment l'interpréter ?**

Une boîte de dispersion permet de représenter schématiquement les principales caractéristiques d'une distribution, notamment en utilisant les quartiles. Son intérêt réside donc principalement dans une représentation graphique d'un caractère quantitatif pour comparer visuellement des séries statistiques.
Pour l'interpréter : la boîte contient 50% des données [Q1 - Q3], le trait au centre correspond à la médiane, et les moustaches ou segments aux extrémités donnent la valeur maximale et minimale donc la dispersion totale.

6. Paramètres de forme

- **Quelle différence faites-vous entre les moments centrés et les moments absolus ? Pourquoi les utiliser ?**

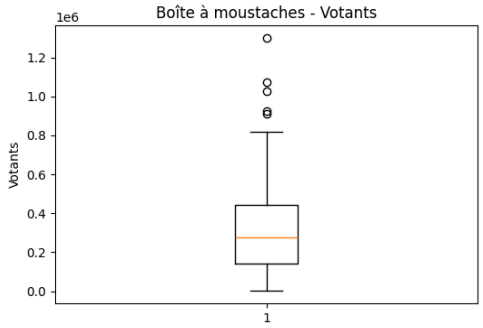
Les moments centrés sont calculés par rapport à la moyenne, tandis que les moments absolus se fondent sur la valeur absolue des écarts à un point donné. Ces derniers sont utiles pour étudier la forme de la distribution ou l'ampleur des écarts.

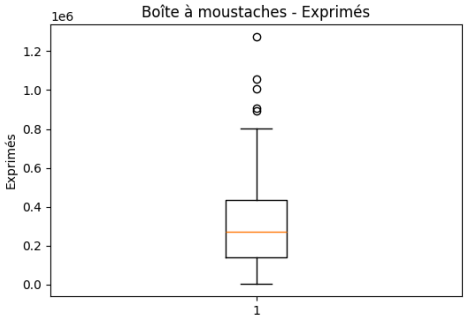
- **Pourquoi vérifier la symétrie d'une distribution et comment faire ?**

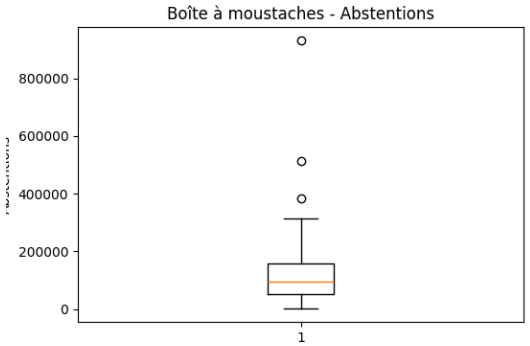
Vérifier la symétrie d'une distribution permet de mettre en lumière la forme de la distribution (pour une distribution symétrique le mode, la moyenne et la médiane sont égaux). Vérifier la symétrie permet d'affirmer ou non si l'on a une dissymétrie, positive ou négative.
Pour cela il faut calculer le coefficient d'asymétrie $\beta 1$, s'il est supérieur à 0 on a une distribution étalée sur la droite, s'il est inférieur à 0 elle est étalée sur la gauche, et s'il est égale à 0 la distribution est symétrique.

II- Mise en oeuvre avec Python

8. À l'aide de Matplotlib et d'une boucle, faire des boîtes à moustache de chaque colonne quantitative. Stocker les résultats dans un dossier *img*







Les boîtes à moustaches des votants, des exprimés et des abstentions présentent des formes assez proches. Dans les trois cas, la distribution est nettement dissymétrique à droite, comme l'indique la position basse de la médiane dans la boîte et la présence de nombreuses valeurs aberrantes élevées. Les ensembles « votants » et « exprimés » sont presque superposables, tant par leur médiane que par leur dispersion, du fait entre les deux variables. Cela révèle aussi une concentration électorale dans un nombre limité de territoires. La distribution des abstentions se distingue par une médiane plus faible et une dispersion légèrement réduite.

Séance 4 : Les distributions statistiques

I- Questions de cours

1. Quels critères mettriez-vous en avant pour choisir entre une distribution statistique avec des variables discrètes et une distribution statistique avec des variables continues ?

Un des critères le plus important pour choisir entre une distribution statistique avec des variables discrètes ou avec des variables continues, est la nature du phénomène et des données. Une variable discrète prend des valeurs finies ou dénombrables. Une variable continue peut prendre toutes les valeurs d'un intervalle réel.

2. Expliquez selon vous quelles sont les lois les plus utilisées en géographie ?

Les lois les plus utilisées en géographie sont la loi de Zipf, notamment pour relier le rang d'une ville à sa population au sein d'un territoire. Mais aussi, la loi log-normale et la loi normale même si cette dernière n'est pas toujours adaptée.

II- Mise en oeuvre avec Python

2. Faire des fonctions (informatiques) pour calculer la moyenne et l'écart type des distributions précédentes.

Loi de Dirac | moyenne = 3.0000 | écart type = 0.0000
Loi Uniforme discrète | moyenne = 3.5000 | écart type = 1.7078
Loi Binomiale | moyenne = 5.0000 | écart type = 1.5811
Loi de Poisson (discrète) | moyenne = 3.0000 | écart type = 1.7321
Loi Zipf-Mandelbrot | moyenne = inf | écart type = inf
Loi Normale | moyenne = 0.0000 | écart type = 1.0000
Loi Log-normale | moyenne = 1.1331 | écart type = 0.6039
Loi Uniforme continue | moyenne = 0.5000 | écart type = 0.2887
Loi Chi-deux | moyenne = 4.0000 | écart type = 2.8284
Loi Pareto | moyenne = 1.5000 | écart type = 0.8660

Voici le résultat obtenu pour le calcul des moyennes et écarts-types de chaque distribution statistique (continues ou discrètes). On peut voir que selon la nature de la loi et de la distribution, le calcul de la moyenne et de l'écart-type diffèrent fortement.

Séance 5 : Les statistiques inférentielles

I- Questions de cours

1. Comment définir l'échantillonnage. Pourquoi ne pas utiliser la population en entier? Quelles sont les méthodes d'échantillonnage ? Comment les choisir ?

- L'échantillonnage consiste à prélever au hasard une partie d'une population mère, de taille fixée n, afin d'en déduire des informations sur la population totale.
- L'étude d'une population en entier est souvent impossible ou trop coûteuse, soit en raison de la taille de la population, soit pour des raisons matérielles et financières. On recourt donc à un échantillon.
- On distingue 3 méthodes d'échantillonnage : les méthodes aléatoires (sondage aléatoire simple, tirage avec remise ou sans remise, échantillons indépendants ou appariés), les méthodes non aléatoires (échantillonnage systématique), et les méthodes Monte-Carlo (lorsque les calculs analytiques sont complexes ou impossibles).
- Pour les choisir, le critère central est la représentativité. Un petit échantillon représentatif est préférable à un grand échantillon biaisé. Cela dépend aussi des contraintes pratiques, de l'existence d'une base de sondage et du coût des tirages

2. Comment définir un estimateur et une estimation ?

Un estimateur est une variable aléatoire, fonction des données observées, construite pour approcher la valeur inconnue d'un paramètre de la population

Une estimation est la valeur numérique obtenue lorsque l'estimateur est calculé à partir des observations effectives de l'échantillon.

3. Comment distingueriez-vous l'intervalle de fluctuation et l'intervalle de confiance ?

L'intervalle de fluctuation suppose que la proportion théorique de la population est connue. Il sert à vérifier si une fréquence observée est compatible avec cette valeur théorique, dans un cadre d'échantillonnage.

L'intervalle de confiance est utilisé lorsque le paramètre est inconnu. Il relie un estimateur à la valeur réelle du paramètre et exprime l'incertitude de l'estimation

L'un relève de la décision, l'autre de l'estimation.

4. Qu'est-ce qu'un biais dans la théorie de l'estimation ?

Un biais est la différence entre l'espérance de l'estimateur et la valeur réelle du paramètre. Un estimateur est sans biais si cette différence est nulle, sinon, il produit une erreur systématique

5. Comment appelle-t-on une statistique travaillant sur la population totale ? Faites le lien avec la notion de données massives ?

Une statistique travaillant sur toute la population est une statistique exhaustive. Elle contient toute l'information disponible sur le paramètre étudié.

Le lien avec les données massives est direct : lorsque l'on dispose de très grandes bases de données proches de l'exhaustivité, l'approche statistique se rapproche davantage du recensement que de l'inférence.

6. Quels sont les enjeux autour du choix d'un estimateur ?
Le choix d'un estimateur relève de plusieurs enjeux, il vise notamment à minimiser le biais, réduire la variance, garantir la convergence et conserver le maximum d'information contenue dans les données. Un mauvais estimateur peut produire des résultats systématiquement erronés.

7. Quelles sont les méthodes d'estimation d'un paramètre ? Comment en sélectionner une ?
Les méthodes d'estimation d'un paramètre la recherche d'estimateurs sans biais et convergents, l'utilisation de statistiques exhaustives, l'analyse de l'information de Fisher, la minimisation de l'erreur quadratique moyenne
Le choix dépend de la loi sous-jacente, de la précision attendue et de la robustesse souhaitée.

8. Quels sont les tests statistiques existants ? À quoi servent-ils ? Comment créer un test ?
Les tests statistiques servent à prendre une décision sous risque d'erreur, par exemple accepter ou rejeter une hypothèse concernant un paramètre inconnu. Pour créer un test, il faut : définir une hypothèse nulle, choisir une statistique de test, fixer un seuil de risque, comparer la statistique observée à une loi de référence.

II- Mise en oeuvre avec python

Résultat sur le calcul d'un intervalle de fluctuation
Fréquences population mère : {'Pour': 0.39, 'Contre': 0.42, 'Sans opinion': 0.19}
Fréquences échantillons : {'Pour': 0.39, 'Contre': 0.42, 'Sans opinion': 0.19}
Intervalles de fluctuation : {'Pour': (0.3598, 0.4202), 'Contre': (0.3894, 0.4506), 'Sans opinion': (0.1657, 0.2143)}
Résultat sur le calcul d'un intervalle de confiance
Fréquences observées : [0.395, 0.396, 0.209]
Intervalles de confiance : {'Pour': (0.3647, 0.4253), 'Contre': (0.3657, 0.4263), 'Sans opinion': (0.1838, 0.2342)}
Théorie de la décision
Test 1 : statistique = 0.9639482021309311 p-value = 6.286744082090187e-22
Test 2 : statistique = 0.2608882349902276 p-value = 7.04938990116743e-67
Loi-normale-Test-1 ne suit pas une loi normale
Loi-normale-Test-2 ne suit pas une loi normale