

Chloé MONDON

M1 GAED

GéoSuds

Rapport analyse de données en géographie

Parcours : débutants

Séance 2 - Les principes généraux de la statistique

1/ Questions de cours

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

La géographie a un positionnement ambigu par rapport à la statistique ; ces deux disciplines entretiennent un rapport complexe et paradoxal. D'un côté, la géographie a tendance à négliger, voire à mépriser la science statistique, car elle veut s'imposer comme une discipline indépendante du champ des statistiques. D'un autre côté, elle produit énormément de données qui ne peuvent qu'être analysées avec l'outil statistique.

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

La question de l'existence du hasard en géographie s'inscrit dans le débat entre deux grandes conceptions héritées de la philosophie. La première, celle du déterminisme, affirme que le hasard n'existe pas : tout a une cause (Laplace). La seconde admet au contraire un hasard apparent, lié à des causes cachées à la raison humaine qu'un progrès futur de la connaissance pourrait expliquer un jour. En géographie, cette tension se retrouve entre nécessité et contingence : certains phénomènes restent impossibles à prévoir précisément malgré l'identification de facteurs géographiques. Toutefois, l'usage des statistiques montre que le hasard peut être appréhendé, c'est-à-dire qu'on ne peut pas prévoir chaque action individuelle, mais on peut dégager des tendances globales, ce qui fonde l'approche multiscalaire propre à la géographie. Ainsi, le hasard existe en géographie dans le détail des réalisations, mais il n'empêche pas l'établissement de tendances, permettant ainsi à la géographie de se constituer comme une science, "la science des échelles".

3. Quels sont les types d'information géographique ?

L'information géographique est divisée en deux grands types. Le premier correspond aux données décrivant les caractéristiques d'un territoire qui peuvent relever soit de la géographie humaine soit de la géographie physique. Dans un système d'information géographique (S.I.G.), ces données constituent la base attributaire, c'est-à-dire les attributs associés à chaque unité spatiale. Le second type d'information concerne la morphologie même des ensembles délimités ; cela constitue les données géométriques du S.I.G.

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

La géographie, au niveau de l'analyse de données, a besoin de visualiser les données. Pour cela, elle doit les collecter et les structurer. Elle a également besoin de comprendre les relations spatiales entre variables afin de faire des comparaisons multiscalaires et/ou de prévoir des évolutions.

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

La statistique descriptive vise à décrire une ou plusieurs variables à l'aide de paramètres de position, de dispersion et de forme. Elle s'appuie sur des tableaux et des graphiques dans le but de synthétiser les données, de les visualiser et de préparer des comparaisons. Elle a pour objectif de décrire et visualiser les données. La statistique explicative, quant à elle, a pour objectif d'expliquer ou de prédire à partir d'un modèle. Donc, elle analyse les relations entre une variable à expliquer et une ou plusieurs variables explicatives dans le but de tester des hypothèses.

6. Quels sont les types de visualisation de données en géographie ? Comment choisir ceux-ci ?

Il existe plusieurs types de visualisation des données en géographie. Par exemple, on trouve les diagrammes en bâtons, les diagrammes circulaires, les histogrammes, les courbes cumulatives et les cartes statistiques (anamorphose, carte de points etc.). Le choix de ceux-ci peut dépendre du type de variable (si elle est qualitative ou quantitative), de l'objectif de l'analyse (comparaison, évolution, ou distribution) et du niveau d'agrégation spatiale.

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

Les méthodes d'analyse de données peuvent être descriptives, explicatives ou de prévision. Les méthodes descriptives permettent de visualiser les données. Les méthodes explicatives relient une variable à expliquer à des variables explicatives. Enfin, les méthodes de prévision servent à modéliser les séries chronologiques.

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

- La population statistique correspond à un ensemble d'unités au sens mathématique du terme.
- L'individu statistique correspond à un élément de la population statistique étudiée. C'est une unité spatiale que l'on peut localiser et cartographier.
- Les caractères statistiques sont des propriétés que l'on observe sur chaque individu (comme l'âge, le revenu etc).
- Les modalités statistiques sont des valeurs possibles prises par un caractère.
- Les deux types de caractères sont qualitatifs et quantitatifs. Les caractères qualitatifs peuvent être nominaux ou ordinaux. Les caractères quantitatifs peuvent être discrets ou continus. Il existe une hiérarchie ; les caractères

9. Comment mesurer une amplitude et une densité ?

Pour mesurer l'amplitude, il faut calculer la longueur du segment qui définit une classe, donc il faut soustraire la valeur minimale a de la classe à sa valeur maximale b . L'amplitude concerne toujours une classe précise.

Pour mesurer la densité, on rapporte l'effectif de la classe (noté n_i) à l'amplitude de cette même classe. Le d signifie densité :

$$d = \frac{n_i}{b - a}$$

10. A quoi servent les formules de Sturges et de Yule ?

Ces formules permettent de déterminer le nombre optimal de classes lors de la discrétisation d'une variable quantitative. Elles évitent la perte d'information en permettant un découpage ni trop fin ni trop grossier.

11. Comment définir un effectif ? Comment calculer une fréquence et une fréquence cumulée ? Qu'est-ce qu'une distribution statistique ?

- L'effectif, aussi appelé fréquence absolue, correspond au nombre d'apparition de la variable.
- La fréquence relative se calcule en faisant le rapport entre l'effectif et l'effectif total. Alors que la fréquence cumulée s'obtient en faisant la somme des fréquences associées aux valeurs inférieures ou égales à k . On peut aussi la calculer en divisant l'effectif cumulé par l'effectif total.

$$\sum_{i=1}^k f_i = \sum_{i=1}^k \frac{n_i}{n} = \frac{1}{n} \sum_{i=1}^k n_i \leq k$$

- La distribution statistique représente la répartition des valeurs d'un caractère dans la population.

Séance 3 - Les paramètres statistiques élémentaires

1/ Questions de cours

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

Le caractère quantitatif est le plus général, car la majorité des paramètres statistiques (moyenne, écart type, variance etc.) sont définis principalement pour des variables quantitatives, et seulement de façon ponctuelle pour des variables qualitatives. Les outils de mesure de position, de dispersion et de forme reposent sur des opérations numériques, ce qui nécessite une variable mesurable.

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

Les caractères quantitatifs discrets prennent un nombre fini ou dénombrable de valeurs distinctes. Ils sont décrits à l'aide de sommes et de fréquences. Alors que les caractères quantitatifs continus peuvent prendre une infinité de valeurs dans un intervalle. Leur étude repose sur des densités de probabilité et des intégrales. Il est nécessaire de les distinguer car les formules statistiques diffèrent : une somme est utilisée pour les variables discrètes, tandis qu'une intégrale est employée pour les variables continues (par exemple pour le calcul de la moyenne).

3. Paramètres de position

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

Il existe plusieurs types de moyenne car chaque moyenne répond à des propriétés et à des usages spécifiques selon la nature des données. Par exemple, la moyenne arithmétique est sensible aux valeurs extrêmes, tandis que d'autres moyennes (harmonique, géométrique, quadratique) sont adaptées à des contextes particuliers (vitesses, produits, puissances). Le choix dépend donc du phénomène étudié.

❖ Pourquoi calculer une médiane ?

La médiane permet de partager une population en deux parties de même effectif. Contrairement à la moyenne arithmétique, elle n'est pas influencée par les valeurs extrêmes. Elle est particulièrement utile pour résumer des distributions dissymétriques.

❖ Quand est-il possible de calculer un mode ?

Le mode peut être calculé lorsqu'une modalité présente un effectif maximal ou une densité maximale. Il ne peut pas exister ou ne pas être unique. En cas de plusieurs modes, la distribution est dite bimodale ou plurimodale, ce qui peut indiquer la présence de plusieurs populations distinctes.

4. Paramètres de concentration

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

La médiale permet de partager la masse totale d'une variable en deux parties égales, et non les effectifs. Elle est pertinente pour analyser la concentration d'une variable (par exemple les revenus). L'indice de concentration de C.Gini compare la médiale à la médiane et mesure l'intensité de la concentration. Plus l'écart entre médiale et médiane est important par rapport à l'étendue, plus la concentration est forte. Il permet ainsi d'évaluer le caractère égalitaire ou inégalitaire 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 ?

L'écart à la moyenne simple n'est pas exploitable car les écarts positifs et négatifs se compensent. La variance utilise le carré des écarts, ce qui permet de mesurer efficacement la dispersion. L'écart type est ensuite utilisé car il s'exprime dans la même unité que la moyenne, ce qui facilite l'interprétation.

❖ Pourquoi calculer l'étendue ?

Calculer l'étendue permet de mesurer rapidement l'intervalle de variation d'une série à partir des valeurs extrêmes. Elle est simple à calculer, mais peu robuste car elle ne tient compte que des valeurs minimale et maximale.

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

Un quantile sert à diviser une série ordonnée en parts égales afin d'analyser la répartition des données. Les quantiles les plus utilisés sont les quartiles, notamment Q1, Q2 (la médiane) et Q3, qui permettent d'étudier la dispersion centrale.

❖ Pourquoi construire une boîte de dispersion ? Comment l'interpréter ?

La boîte de dispersion permet de représenter graphiquement les principales caractéristiques d'une distribution (quartiles, médiane, valeurs extrêmes). Elle sert à comparer visuellement plusieurs séries statistiques et à analyser la dispersion, la position centrale et les éventuelles asymétries.

6. Paramètres de forme

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

Les moments absolus sont calculés à partir des valeurs brutes de la variable, tandis que les moments centrés sont calculés par rapport à la moyenne. Les moments centrés permettent de caractériser la dispersion, la dissymétrie et l'aplatissement d'une distribution. Ils sont essentiels pour décrire la forme globale de la distribution.

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

Vérifier la symétrie permet de mieux comprendre la structure d'une distribution et de savoir si la moyenne, la médiane et le mode coïncident. La symétrie est analysée à l'aide du coefficient d'asymétrie β_1 :

- $\beta_1 > 0$: dissymétrie positive
- $\beta_1 < 0$: dissymétrie négative
- $\beta_1 = 0$: distribution symétrique

Séance 4 - Les distributions statistiques

1/ 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 ?

Le choix entre une distribution statistique à variables discrètes et une distribution à variables continues dépend à la fois de la nature des données analysées et des objectifs de l'étude. Les distributions à variables discrètes sont adaptées aux données dénombrables, correspondant à des valeurs entières. Elles permettent d'analyser les fréquences et la répétition de valeurs ou de quantités comptables. A l'inverse, les distributions à variables continues conviennent aux données mesurables sur un intervalle, généralement décimal, et pouvant s'étendre sur une échelle théoriquement infinie, comme la température d'une zone donnée. Elles sont donc privilégiées pour l'étude de valeurs appartenant à des intervalles continus, éventuellement infinis.

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

Au regard des caractéristiques propres à chaque distribution statistique et des besoins spécifiques de la géographie, certaines lois sont particulièrement mobilisées dans cette discipline.

- La loi normale est l'une des distributions les plus couramment utilisées. Elle repose sur une variable aléatoire influencée par de nombreux facteurs extérieurs et indépendants. En s'appuyant sur le théorème central limite, qui considère la somme des multiples effets de faible amplitude, la loi normale offre une représentation pertinente des facteurs intervenant dans des phénomènes physiques, spatiaux ou sociaux.
- La loi uniforme continue décrit une répartition équitable des valeurs d'une variable sur un intervalle donné. Elle est notamment utilisée pour simuler des échantillonnages, des positions ou des modèles en géographie.
- La loi de Poisson permet de modéliser la fréquence d'événements rares au sein d'un grand nombre d'observations, tels que le nombre d'accidents aériens par an ou de séismes par décennie et par continent. En géographie, elle est particulièrement utile

pour analyser la répartition spatio-temporelle de phénomènes peu fréquents ou atypiques.

- La loi de Zipf-Mandelbrot met en évidence une relation inverse entre la fréquence d'une variable et son rang. Elle est utilisée pour étudier les relations rang-taille, notamment dans l'analyse de la population des villes, et permet ainsi de mettre en évidence des hiérarchies urbaines.
- La loi-log-normale s'applique à des variables soumises à des processus multiplicatifs, générant des effets de croissance proportionnelle. Elle permet par exemple de représenter des distributions telles que celle des revenus au sein d'un espace donné.

Séance 5 - Les statistiques inférentielles

1/ 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 correspond à la sélection d'un sous-ensemble issu d'une population mère, c'est-à-dire l'ensemble total des individus concernés par l'étude. Il est généralement préférable de travailler sur un échantillon plutôt que sur la population entière, cette dernière étant souvent trop vaste pour être analysée dans sa globalité. Afin de permettre la généralisation des résultats à l'ensemble de la population mère, l'échantillon doit être représentatif, aléatoire et adapté à l'objet de recherche. Il existe plusieurs méthodes d'échantillonnage :

- Les méthodes aléatoires reposent sur un tirage au sort effectué à partir d'une base de sondage. Les individus, numérotés de 1 à N , sont sélectionnés selon un procédé aléatoire, avec ou sans remise. Dans ce type de méthode, chaque individu possède la même probabilité d'être choisi, garantissant ainsi l'absence de biais. Les méthodes aléatoires sont particulièrement adaptées aux études pour lesquelles la répétition des tirages n'entraîne pas de conséquences majeures.
- Les méthodes non aléatoires visent à constituer un échantillon dit représentatif, considéré comme un "modèle réduit" de la population mère, à partir de procédés de sélection ciblés. Elles incluent notamment l'échantillonnage systématique, qui

consiste à sélectionner des individus à intervalles réguliers dans une base de sondage, ainsi que la méthode des quotas, qui respecte les proportions de certaines caractéristiques de la population (âge, nationalité, genre, etc.). Ces méthodes sont privilégiées lorsque l'objectif est de reproduire fidèlement les principales caractéristiques distinctives de la population mère.

2. Comment définir un estimateur et une estimation ?

- Un estimateur est associé à une variable aléatoire et prend une valeur supposée proche du paramètre réel étudié. Il s'agit d'une fonction des observations permettant d'évaluer les paramètres de la population mère à partir des informations recueillies au sein de l'échantillon.
- L'estimation correspond à une démarche d'inférence statistique visant à fournir des évaluations fiables des caractéristiques de la population mère à partir des données issues d'un échantillon.

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

L'intervalle de fluctuation repose sur la connaissance préalable de la proportion théorique réelle de la population. Il définit, généralement au seuil de 95%, un intervalle dans lequel la proportion observée est censée se situer lorsque la valeur théorique p est connue ou supposée. A l'inverse, l'intervalle de confiance correspond à la marge d'erreur associée à une estimation issue d'un échantillon lorsque la valeur de p est inconnue. Il permet ainsi d'encadrer l'estimation obtenue à partir de l'échantillon et d'évaluer l'incertitude liée à l'estimation des caractéristiques de la population mère.

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

Dans la théorie de l'estimation, le biais correspond à l'écart entre l'espérance d'un estimateur et la valeur réelle du paramètre au sein de la population. Lorsqu'un estimateur est biaisé, il engendre une erreur systématique : ses valeurs tendent à se concentrer autour de son espérance mathématique plutôt qu'autour de la véritable valeur du paramètre à estimer.

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

Une statistique portant sur l'ensemble de la population est qualifiée d'enquête exhaustive. Elle s'oppose au sondage, qui repose sur l'analyse d'un échantillon représentatif de la population mère. De manière comparable, les données massives se distinguent par leur volume et leur diversité particulièrement élevés, ce qui requiert des méthodes d'analyse et des infrastructures technologiques spécifiques, telles que les centres de données. Ainsi, le traitement d'une enquête exhaustive s'apparente à celui des données massives, dans la mesure où il implique l'analyse d'un volume très important et d'une grande variété de données.

6. Quels sont les enjeux autour du choix d'un estimateur ?

Le choix d'un estimateur repose sur plusieurs critères, notamment le biais, l'efficacité et la convergence. Le biais mesure l'écart entre l'espérance de l'estimateur et la valeur réelle du paramètre étudié ; un estimateur est dit sans biais lorsque ces deux valeurs coïncident. L'efficacité d'un estimateur, évaluée par sa variance, est d'autant meilleure que celle-ci est faible, traduisant une moindre dispersion des estimations. Enfin, la convergence d'un estimateur correspond à la tendance de sa distribution à se rapprocher de la valeur réelle du paramètre lorsque la taille de l'échantillon augmente vers l'infini, ce qui constitue un indicateur essentiel de sa fiabilité et de sa représentativité.

7. Quelles sont les méthodes d'estimation d'un paramètre ? Comment en sélectionner une ?

Il existe plusieurs méthodes pour estimer un paramètre. La plus simple, la méthode des moindres carrés, consiste à minimiser la somme des carrés des écarts entre les valeurs observées et les valeurs ajustées, c'est-à-dire les résidus. En revanche, la méthode du maximum de vraisemblance est plus générale et souvent préférable. Elle repose sur l'évaluation de la probabilité des différentes valeurs possibles du paramètre, en se basant sur les observations disponibles. L'objectif de cette méthode est de choisir la valeur du paramètre qui maximise cette vraisemblance, permettant ainsi d'inférer les paramètres de probabilité à partir de l'échantillon étudié.

8. Quels sont les tests statistiques existants ? À quoi servent-ils ? Comment créer un test ?

La statistique inférentielle repose sur différents tests, regroupés au sein de la théorie des tests, qui visent à formuler et à évaluer des hypothèses portant sur les paramètres ou les lois

intervenant dans les phénomènes étudiés. Un test statistique se construit en formulant une hypothèse nulle et une hypothèse alternative, en choisissant un test adapté à la nature des données et à la distribution supposée, puis en comparant la statistique observée à un seuil de décision afin d'accepter ou de rejeter l'hypothèse nulle. La théorie de la décision s'articule ainsi autour de plusieurs types de tests statistiques :

- Les tests paramétriques supposent que la forme de la distribution des données est connue et portent sur un paramètre précis. Ils reposent sur une hypothèse nulle formulée à propos de la loi des données et concernent notamment des paramètres tels que la moyenne, l'écart type ou le type de distribution.
- Les tests non paramétriques ne tiennent pas compte de la forme de la distribution et peuvent s'appliquer aussi bien à des variables qualitatives que quantitatives. Ils portent généralement sur des paramètres comme l'effectif ou la médiane.
- Les tests de signification évaluent l'écart entre la distribution observée et l'hypothèse nulle, à partir de la valeur du score observé lors de l'expérience.
- Les tests robustes, aussi appelés tests libres, sont valables quelle que soit la loi de la variable étudiée. Ils sont utilisés lorsque la distribution de la variable aléatoire n'est pas connue.
- Les tests d'hypothèse visent à déterminer si les données issues de l'échantillon sont compatibles avec l'hypothèse formulée sur la population mère. Ils permettent d'accepter ou de rejeter l'hypothèse nulle, utilisée comme référence lors du test.
- Les tests d'ajustement servent à apprécier l'adéquation entre une situation observée et un modèle théorique, en vérifiant si la loi de probabilité supposée correspond aux réalisations observées d'un échantillon.
- Les tests de comparaison sont employés pour confronter plusieurs échantillons entre eux.
- Les tests d'indépendance interviennent lorsque l'analyse porte sur plusieurs variables aléatoires, afin d'évaluer l'existence ou non d'un lien entre elles.

9. Que pensez-vous des critiques de la statistique inférentielle ?

Les critiques de la statistique inférentielle soulignent que certaines méthodes peuvent conduire à des conclusions trop hâtives concernant la représentativité de l'échantillon par rapport à la population mère ou le rejet de l'hypothèse nulle. Toutefois, malgré ces limites, leur utilité demeure majeure : la statistique inférentielle permet d'analyser les phénomènes de

grande échelle à partir de données d'échantillon, tout en intégrant des outils destinés à encadrer et à maîtriser les marges d'erreur propres à chaque test.

Séance 6 - La statistique d'ordre des variables qualitatives

1/ Questions de cours

- 1. Qu'est-ce qu'une statistique ordinale ? À quel autre statistique catégorielle s'oppose-t-elle ? Quel type de variables utilise-t-elle ? En quoi cela peut matérialiser une hiérarchie spatiale ?**

La statistique ordinale est une approche qui analyse la progression, la stabilité ou le recul d'une entité au sein d'un classement, à partir de variables ordinales organisées selon un ordre croissant, également appelé ordre naturel. Elle permet notamment de représenter des hiérarchies spatiales en mesurant ou en visualisant la position, la fréquence ou l'évolution d'une entité dans un espace donné. A l'inverse, les statistiques nominales portent sur des catégories qui ne suivent aucun ordre hiérarchique particulier.

- 2. Quel ordre est à privilégier dans les classifications ?**

Dans les classifications, l'ordre croissant, également appelé ordre naturel, est à privilégier, car il facilite l'identification des valeurs aberrantes au sein d'une série d'observation et permet également de mettre en évidence la loi de la valeur la plus élevée.

- 3. Quelle est la différence entre une corrélation des rangs et une concordance de classements ?**

Bien que la corrélation des rangs et la concordance de plusieurs classements aient toutes deux pour objectif d'évaluer la similarité entre des classements, elles reposent sur des approches distinctes. La corrélation des rangs mesure le degré de ressemblance entre deux classements ordinaux et permet également d'apprécier la signification statistique du lien entre eux. A l'inverse, la concordance de plusieurs classements vise à généraliser cette analyse à plus de deux classements et à évaluer la cohérence des positions occupées par les mêmes individus au sein de ces différents classements.

4. Quelle est la différence entre les tests de Spearman et de Kendal ?

- Le test de Spearman est utilisé pour analyser la corrélation des rangs et permet de déterminer si deux classements sont similaires, inversés ou indépendants. Il repose sur la formulation de deux hypothèses — soit un coefficient de corrélation non significativement différent de zéro, soit un coefficient significativement différent de zéro. En géographie, ce test peut notamment être mobilisé pour classer des villes selon leur population ou pour comparer plusieurs variables liées à un même objet d'étude, afin d'évaluer la dépendance entre différents classements.
- À l'inverse, le test de Kendall s'applique à deux classements issus de séries de valeurs distinctes en formant des couples de rangs. Il consiste à vérifier si l'ordre naturel des variables est respecté, en distinguant les paires dites concordantes (+1) et discordantes (-1). Après l'examen de l'ensemble des paires, un coefficient Tau égal à +1 indique des classements identiques, tandis qu'un coefficient de -1 révèle des classements inverses.

Ainsi, bien que les tests de Spearman et de Kendall poursuivent tous deux l'objectif d'évaluer la similarité entre deux classements, ils se distinguent par leur méthode de calcul : le test de Spearman mesure directement la corrélation globale entre deux classements, tandis que le test de Kendall s'appuie sur le décompte des paires concordantes et discordantes pour exprimer la relation.

5. À quoi servent les coefficients de Goodman-Krusdal et de Yule ?

Le coefficient Γ de Goodman-Krusdal est utilisé afin de calculer la proportion de surplus de paires concordantes par rapport aux paires discordantes, et varie entre -1 et +1. Le coefficient Q de Yule suit la même logique que le coefficient Γ de Goodman-Krusdal, mais est appliqué aux tableaux de contingence 2 x 2, lui-même calculé à partir des fréquences observées. En créant une table de contingence pour évaluer la fréquence des événements, le coefficient de Yule calcule l'association des séries ordinales entre elles. Celui-ci varie entre -1 (l'association est négative totale) et +1 (l'association est positive parfaite).