

A

Aléa. L'aléa traduit les effets des variations dues au hasard ou les effets accidentels. Cette composante est notée $\varepsilon(t)$.

Analyse en composantes principales. L'analyse en composantes principales (ACP) est une méthode d'analyse multivariée descriptive qui permet de décrire un ensemble d'individus par un ensemble de variables quantitatives.

Analyse en composantes principales normée. Une analyse en composantes principales réalisée sur des données centrées réduites est dite normée.

Axe factoriel. Les grands axes de dispersion des points qui représentent les individus dans l'espace défini par les p variables d'origine sont appelés des axes factoriels.

B

Base de sondage. La liste exhaustive de tous les individus qui composent la population est appelée base de sondage.

Boîte de dispersion. Une boîte de dispersion est une représentation graphique qui permet de visualiser les quartiles, l'étendue et l'intervalle interquartile pour une variable donnée.

C

Cercle de corrélation. Un cercle de corrélation est un plan de projection du nuage des points variables, défini par deux axes factoriels.

Classification. Les méthodes de classification, aussi appelées « typologies », visent à créer des ensembles d'individus, aussi appelés « groupes » ou « classes », ayant des caractéristiques proches pour les variables prises en compte dans l'étude.

Classification hiérarchique. Les méthodes hiérarchiques consistent à effectuer un ensemble de partitions successives emboîtées les unes dans les autres.

Classification mixte. Une classification mixte consiste à combiner les méthodes hiérarchiques et non hiérarchiques.

Classification non hiérarchique. Les méthodes non hiérarchiques regroupent, par itérations successives, les individus en un nombre de classes fixé au départ.

Coefficient de corrélation linéaire. L'indicateur utilisé pour juger de l'intensité de la corrélation linéaire entre deux variables quantitatives X et Y s'appelle le coefficient de corrélation linéaire. Il est noté $\rho_{X,Y}$ dans une population et $r_{X,Y}$ dans un échantillon.

Coefficient de détermination. La qualité de la modélisation par une régression multiple se mesure par le pourcentage de variance de Y expliquée par l'ensemble des variables explicatives, le coefficient de détermination noté ρ^2 dans une population et r^2 dans un échantillon.

Coefficient de variation. Le coefficient de variation est l'écart-type divisé par la moyenne. Il permet de juger si une variable est faiblement ou fortement dispersée.

Coefficient du khi-deux. Le coefficient du khi-deux mesure l'intensité du lien entre deux variables qualitatives X et Y . Il est noté $\chi^2_{X,Y}$ sur la population, $c^2_{X,Y}$ lorsqu'il est calculé sur l'échantillon.

Coefficients de la droite de régression. β_0 est la constante de la droite de régression (ou ordonnée à l'origine) et β_1 la pente de la droite de régression. β_0 et β_1 sont aussi appelés les paramètres ou coefficients du modèle de régression. Les estimations de β_0 et β_1 par la méthode des moindres carrés ordinaires sont notées b_0 et b_1 .

Contribution relative. La contribution relative d'un individu à un axe mesure sa participation à la formation d'un axe. Elle s'exprime en pourcentage.

Corrélation linéaire. Il y a corrélation linéaire en deux variables quantitatives lorsqu'elles sont liées par une relation linéaire.

Courbe d'ajustement. Une courbe qui ajuste un phénomène observé s'appelle courbe d'ajustement. À cette courbe est associée une série de valeurs, dite série ajustée, notée \hat{y} .

D

Décile. Les déciles sont les valeurs qui partagent la distribution ordonnée en 10 classes de même effectif.

Distribution conditionnelle. Dans le cas de la distribution conjointe de deux variables qualitatives X et Y , respectivement à p et q modalités, on appelle distribution conditionnelle de X toute distribution statistique de X observée sur une sous-population définie par une modalité de la variable Y .

Distribution d'échantillonnage. La distribution des valeurs prises par un estimateur sur l'ensemble des échantillons de même

taille issus de la population est appelée distribution d'échantillonnage de l'estimateur.

Distribution marginale. Dans le cas de la distribution conjointe de deux variables qualitatives X et Y , la distribution marginale de X (respectivement Y) est la distribution de X (respectivement Y) étudiée indépendamment de l'observation de Y (respectivement X).

Droite des moindres carrés. La droite obtenue par la méthode des moindres carrés s'appelle la droite des moindres carrés ou droite de régression. Elle représente la relation linéaire entre deux variables, X et Y .

E

Écart-type. L'écart-type d'une variable X est la racine carrée de la variance. Il est noté σ_X dans la population et s_X dans l'échantillon.

Écart-type corrigé. L'écart-type corrigé d'une variable X est la racine carrée de la variance corrigée. Il est noté \bar{s}_X .

Échantillon. Un échantillon est un groupe d'individus extrait de la population.

Échantillon aléatoire. Un échantillon aléatoire est constitué par un mécanisme aléatoire qui respecte la probabilité connue et non nulle, pour chaque individu, d'appartenir à cet échantillon.

Échantillons appariés. Deux échantillons sont dits appariés lorsque les individus sont les mêmes dans les deux échantillons.

Échantillons indépendants. Deux échantillons sont considérés comme indépendants s'ils ont été tirés indépendamment l'un de l'autre.

Effectif. Le nombre d'individus considérés est l'effectif.

Erreur d'ajustement. L'erreur d'ajustement, appelée aussi résidu ou écart résiduel, est la

différence entre la valeur de Y observée, y_t , et la valeur ajustée par le modèle, \hat{y}_t . Dans le cas d'une série chronologique, l'erreur d'ajustement est la différence entre série observée et série ajustée. Elle est notée $e(t) = y(t) - \hat{y}(t)$.

Erreur d'échantillonnage. Une erreur d'échantillonnage résulte des fluctuations dues au principe même de l'échantillonnage.

Erreur de première espèce. L'erreur de première espèce consiste à rejeter l'hypothèse nulle, alors que celle-ci est vraie.

Erreur de deuxième espèce. L'erreur de deuxième espèce consiste à accepter l'hypothèse nulle, alors que celle-ci est fautive.

Erreur de couverture. L'erreur de couverture provient d'une différence entre la population cible à étudier et la population réellement étudiée.

Erreur de mesure. Un écart entre les réponses enregistrées et les vraies valeurs s'appelle une erreur de mesure.

Erreur de modélisation. L'erreur de modélisation, aussi appelée erreur d'ajustement, provient de l'écart entre le modèle et la réalité.

Erreur de non-réponse. L'erreur de non-réponse provient de l'absence partielle ou complète d'informations concernant les individus de l'échantillon.

Erreur-type. L'écart-type d'un estimateur est aussi appelé erreur-type.

Estimateur. Un estimateur permet de fournir, à partir d'un échantillon, une valeur à un paramètre inconnu caractérisant la population. C'est une variable aléatoire. Dans l'ouvrage, l'estimateur est noté par une majuscule romaine (par exemple, M_X est l'estimateur de la moyenne), la valeur de l'estimateur sur un échantillon par une

minuscule (par exemple, m_X), et une valeur sur la population par une minuscule grecque (par exemple, μ_X est la moyenne dans la population).

Estimateur convergent. Un estimateur convergent donne une valeur qui se rapproche de la vraie valeur du paramètre dans la population, à mesure que la taille de l'échantillon croît.

Estimateur efficace. L'estimateur le plus efficace est celui pour lequel la dispersion autour de la vraie valeur est la plus faible.

Estimateur sans biais. Un estimateur est sans biais si sa distribution d'échantillonnage est centrée autour de la vraie valeur dans la population.

Estimation ponctuelle. La valeur numérique prise par l'estimateur sur l'échantillon dont on dispose s'appelle l'estimation ponctuelle.

Estimer. Estimer consiste, à partir des observations obtenues sur un échantillon, à attribuer des valeurs numériques aux paramètres de la population dont cet échantillon est issu.

Étendue. L'étendue est la différence entre la plus grande et la plus petite valeur prise par la variable.

G

Graphique de la série brute. Le graphique de la série brute représente les valeurs d'une série chronologique $y(t)$ en fonction du temps, t .

Graphique des résidus. Le graphique des résidus croise en abscisse la variable explicative X et en ordonnée les résidus.

Graphique superposé. Dans le graphique superposé, les valeurs d'une série chronologique $y(t)$ sont représentées en superposant

chaque période (par exemple, année ou trimestre).

H

Hypothèse alternative. Dans les tests statistiques, l'hypothèse alternative notée H_1 , exprime un écart, la présence d'un effet, une évolution par rapport à une situation de référence.

Hypothèse nulle. Dans les tests statistiques, l'hypothèse nulle, notée H_0 , exprime une situation de référence, la non-évolution, l'absence d'effet.

Hypothèse statistique. Une hypothèse statistique est une proposition concernant une caractéristique d'une variable sur une population.

I

Individu. Un individu est une unité de la population.

Individu ou variable actif. Les variables et les individus qui participent aux calculs d'une analyse statistique (par exemple dans une ACP, une classification, etc.) sont dits actifs.

Individu ou variable illustratif (ou supplémentaire). Plus particulièrement dans le cas de l'ACP, ces individus ou variables sont intégrés dans l'analyse tout en ne contribuant pas à la formation des axes.

Inertie interclasse (ou intergroupe). L'inertie interclasse (ou intergroupe) mesure l'inertie entre les groupes. Elle est égale à la somme des écarts au carré entre chaque centre de gravité de classe et le centre de gravité du nuage.

Inertie intraclasse (ou intragroupe). L'inertie intraclasse (ou intragroupe) est la somme des inerties de chaque classe.

Inertie totale. La dispersion des points du nuage autour du centre de gravité est mesurée par l'inertie totale. Elle est mesurée par la somme des distances au carré entre chaque individu et le centre du nuage de points.

Intervalle de confiance. Un intervalle de confiance est une fourchette de valeurs qui a une certaine probabilité, appelée niveau de confiance, de contenir la valeur du paramètre sur la population.

Intervalle interquartile. L'intervalle interquartile mesure l'écart entre les valeurs du premier et du troisième quartile.

M

MAD (*Mean Absolute Deviation*). Le MAD (*Mean Absolute Deviation*) est la moyenne des erreurs en valeur absolue.

MAPE (*Mean Absolute Percentage Error*). Le MAPE (*Mean Absolute Percentage Error*) est le pourcentage moyen d'erreurs (prises en valeur absolue).

Marge d'erreur. La marge d'erreur associée à l'estimation trouvée sur l'échantillon est aussi appelée demi-intervalle de confiance.

Médiane. La médiane est la valeur du caractère qui partage une distribution en deux sous-ensembles de même effectif.

Méthode d'extrapolation. Les méthodes d'extrapolation consistent à regarder la forme d'un phénomène observé dans le passé, puis à la projeter dans le futur.

Méthode des quotas. Un échantillon construit par la méthode des quotas est un échantillon qui respecte la répartition de certaines caractéristiques (par exemple, sexe, taille du foyer, etc.) au sein de la population.

Méthode empirique. Dans le cas des méthodes empiriques, la sélection des

données n'est pas effectuée par sélection aléatoire, mais par un choix raisonné.

Méthode explicative. Les méthodes explicatives s'attachent à déterminer une fonction f qui modélise la relation liant p variables explicatives X_1, X_2, \dots, X_p et Y , la variable à expliquer, avec $Y = f(X_1, X_2, \dots, X_p)$.

Mode. Le mode est la modalité ou la valeur qui correspond au plus grand effectif.

Modèle additif. Dans le cas des séries chronologiques, la formule de décomposition d'un modèle additif est notée $y(t) = T(t) + S(t) + \varepsilon(t)$, où $y(t)$ est la série observée au temps t , $T(t)$ la composante tendancielle au temps t , $S(t)$ la composante saisonnière au temps t , et $\varepsilon(t)$ l'aléa au temps t .

Modèle de régression linéaire simple. Le modèle de régression linéaire simple, ou modèle linéaire, est le modèle où la relation entre X et Y est représenté par une droite. L'équation de cette droite est la suivante : $Y = \beta_0 + \beta_1 X + \varepsilon$.

Modèle multiplicatif. Dans le cas des séries chronologiques, la formule de décomposition d'un modèle multiplicatif est notée $y(t) = T(t) \cdot S(t) \cdot \varepsilon(t)$, où $y(t)$ est la série observée au temps t , $T(t)$ la composante tendancielle au temps t , $S(t)$ la composante saisonnière au temps t , et $\varepsilon(t)$ l'aléa au temps t .

Modèle statistique. Un modèle statistique est une simplification de la réalité qui vise à quantifier des relations entre plusieurs variables.

Modéliser. Modéliser consiste à formaliser les relations entre des variables.

Moyenne. La moyenne s'obtient à partir des données brutes en divisant la somme des valeurs observées par l'effectif total. Elle

est notée μ_X dans la population et m_X dans l'échantillon.

Moyenne conditionnelle. Dans le cadre de la distribution d'une variable quantitative Y selon une variable qualitative X , la moyenne de la variable Y calculée pour une des modalités de la variable X est appelée moyenne conditionnelle de Y selon la modalité x_i de X . Elle est notée $\mu_{Y/X=x_i}$.

Moyenne marginale. Dans le cas d'une distribution conjointe, la moyenne d'une variable est appelée moyenne marginale.

Moyenne mobile. La moyenne mobile d'ordre p , $MM_p(t)$, est une série chronologique résultant de la moyenne de p valeurs de la série observée $y(t)$.

MSE (Mean Square Error). Le MSE (*Mean Square Error*) est la moyenne de l'erreur d'ajustement au carré.

Multicolinéarité. Il y a multicolinéarité lorsque des variables explicatives sont fortement corrélées linéairement entre elles.

N

Nuage de points. Un nuage de points consiste en un graphique représentant chaque individu par un point dans le plan défini par deux variables quantitatives.

O

Observation aberrante. Une observation est dite aberrante si elle est atypique, dans le sens où elle s'éloigne des autres observations.

Ordonnée à l'origine. Voir coefficients de la droite de régression.

P

Parangon. Dans une classification, les individus les plus représentatifs de chaque classe sont appelés parangons.

Pente. Voir coefficients de la droite de régression.

Plan factoriel. Un plan factoriel est un plan de projection du nuage des points-individus défini par deux axes factoriels.

Population. La population est constituée de l'ensemble (ou univers) des individus objets de l'étude.

Pourcentage d'inertie. La qualité de représentation d'un nuage par un axe se mesure avec le pourcentage d'inertie (ou de variance) expliquée par l'axe.

Premier axe factoriel. Le premier axe factoriel Δ_1 est défini tel que la dispersion globale des points du nuage dans la direction de Δ_1 soit maximale.

Probabilité critique. On appelle probabilité critique, notée *P* valeur, la plus petite valeur du risque d'erreur de 1^{re} espèce pour laquelle la décision serait de rejeter l'hypothèse nulle.

Proportion. Une proportion est le nombre d'individus suivant une caractéristique divisée par le nombre total d'individus. Elle est notée π_X dans la population et p_X dans l'échantillon.

Puissance du test. La puissance du test est la probabilité d'accepter l'hypothèse nulle, alors qu'elle est vraie. Elle est le complémentaire du risque de 2^e espèce et est notée $1 - \beta$.

Q

Quartile. Les quartiles sont les valeurs du caractère qui partagent la distribution ordonnée en quatre sous-ensembles de même effectif.

R

Rapport de corrélation. Le rapport de corrélation de Y selon X mesure l'intensité du lien entre une variable quantitative Y et une variable qualitative X .

Région critique. Voir zone de rejet

Régression multiple. La régression est dite multiple lorsque la variable à expliquer, notée Y , de type quantitatif, est mise en relation avec p variables explicatives, notées X_1, X_2, \dots, X_p , ($p > 1$).

Résultat significatif. Un résultat est dit significatif s'il a fait l'objet d'un test statistique qui aboutit à rejeter l'hypothèse nulle pour un risque d'erreur de 1^{re} espèce donné.

Risque d'erreur de 1^{re} espèce. Le risque d'erreur de 1^{re} espèce, noté α , est la probabilité de rejeter l'hypothèse nulle, alors qu'elle est vraie.

Risque d'erreur de 2^e espèce. Le risque d'erreur de 2^e espèce, noté β , est la probabilité d'accepter l'hypothèse nulle, alors qu'elle est fautive.

RMSE (Root Mean Square Error). La racine carrée du MSE est nommée RMSE (*Root Mean Square Error*).

S

Saisonnalité. La composante saisonnière, ou saisonnalité, traduit les fluctuations revenant à intervalles réguliers. Elle est notée $S(t)$.

Série chronologique. Une série chronologique (appelée aussi chronique ou série temporelle) est une suite d'observations d'un phénomène dans le temps. Elle est notée $y(t)$.

Série corrigée des variations saisonnières. La série corrigée des variations saisonnières,

notée $CVS(t)$, est la série que l'on obtient une fois que la série observée a été désaisonnalisée.

Sondage aléatoire. L'échantillon est constitué selon un principe aléatoire.

Sondage aléatoire stratifié. Dans un sondage aléatoire stratifié, la population est découpée en plusieurs groupes, appelés strates, puis un tirage aléatoire simple est réalisé dans chacune de ces strates.

Statistique. La statistique est un ensemble de méthodes scientifiques dont l'objectif est d'analyser, structurer et modéliser des informations numériques.

Statistique descriptive. La statistique descriptive a pour objet de résumer et de présenter l'information contenue dans des données collectées sur un groupe d'individus.

Statistique descriptive bivariée. La statistique descriptive bivariée a pour objet d'étudier conjointement deux variables X et Y sur une même population.

Statistique descriptive univariée. La statistique descriptive univariée fournit les outils statistiques pour organiser, présenter et synthétiser l'information issue de l'analyse d'une variable indépendamment des autres.

Statistique du test. Dans un test d'hypothèse, la statistique du test est une variable aléatoire utilisée pour contrôler l'hypothèse nulle.

Statistique inférentielle. La statistique inférentielle consiste à décrire la population à partir d'observations faites sur l'échantillon. Les caractéristiques inconnues d'une population sont déduites à partir d'un échantillon issu de cette population.

Statistique multivariée. La statistique multivariée vise à étudier plusieurs variables

simultanément. Elle peut être de nature descriptive ou explicative.

T

Tableau de contingence. Un tableau de contingence croise les distributions de deux variables. Il est aussi appelé tri croisé.

Tableau des profils-colonnes. Le tableau des profils-colonnes donne les fréquences conditionnelles en colonne.

Tableau des profils-lignes. Le tableau des profils-lignes fournit les fréquences conditionnelles en ligne.

Tableau de distribution. Le tableau de distribution présente, pour chaque modalité ou valeur de la variable, le nombre d'individus (effectif) qui prennent cette modalité ou cette valeur.

Tableau individus-variables. Le tableau individus-variables reporte les valeurs ou les modalités prises par les N individus pour les p variables de l'étude.

Taille de la population ou de l'échantillon. Il s'agit de l'effectif total de l'étude. Dans la population, le nombre d'individus est noté N , dans l'échantillon il est noté n .

Tendance. La composante tendancielle, ou tendance, traduit l'aspect général de la série. Elle est notée $T(t)$.

Test bilatéral. Un test bilatéral est un test où l'hypothèse alternative H_1 se traduit par une différence (\neq).

Test d'association. Les tests d'association visent à vérifier, à partir de données d'échantillon, si des variables sont liées dans une population.

Test d'hypothèse. Les tests d'hypothèse (ou tests statistiques) sont un ensemble de méthodes statistiques qui permettent, à

partir de données d'échantillon, d'accepter ou de rejeter une hypothèse concernant la population d'où est tiré l'échantillon.

Test de comparaison à une norme. Dans le test de comparaison à une norme, aussi appelé test de conformité, l'objectif est de déterminer si un paramètre (une moyenne, une proportion, une variance, etc.) dans une population est égal, supérieur ou inférieur à une norme.

Test de comparaison sur échantillons de deux populations. Dans un test de comparaison sur échantillons de deux populations, l'objectif est de comparer un paramètre (moyenne, proportion, variance, etc.) dans une population avec le même paramètre calculé dans une autre population.

Test de Fisher. Le test de Fisher consiste à s'assurer de la validité globale d'une régression multiple, c'est-à-dire à vérifier que l'ensemble des coefficients $\beta_1, \beta_2, \dots, \beta_p$ sur la population ne sont pas tous nuls simultanément.

Test de Student. Dans une régression, le test de Student est un test de comparaison de la pente à une norme égale à 0 qui permet de conclure sur la validité de la régression simple sur la population.

Test du khi-deux d'indépendance. Le test du khi-deux d'indépendance permet de décider si deux variables qualitatives sont indépendantes sur la population ou, au contraire, liées.

Test unilatéral. Un test unilatéral inclut dans l'hypothèse alternative H_1 un symbole d'inégalité, $<$ ou $>$.

V

Valeur ajustée. La valeur ajustée est la valeur de Y que l'on aurait dû observer si la relation

suivait parfaitement le modèle postulé (par exemple, un modèle linéaire).

Valeur critique. La valeur de la statistique qui sépare zone de non-rejet et zone de rejet s'appelle la valeur critique.

Variable muette. Une variable muette, aussi appelée indicatrice, est une variable prenant deux valeurs 1 ou 0 selon que l'observation a le caractère étudié ($= 1$) ou pas ($= 0$).

Variable statistique. Une variable statistique décrit une caractéristique des individus sur lesquels porte l'étude.

Variable statistique quantitative. Une variable statistique quantitative est une variable associée à un caractère mesurable.

Variable quantitative continue. Une variable quantitative continue est une variable quantitative qui peut prendre toutes les valeurs dans un intervalle donné.

Variable quantitative discrète. Une variable quantitative discrète prend un nombre limité de valeurs entières.

Variable statistique qualitative. Une variable statistique qualitative est une variable associée à un caractère qui n'est pas mesurable.

Variable statistique qualitative nominale. Une variable statistique qualitative nominale est une variable qualitative dont les modalités ne peuvent pas être classées selon un ordre préétabli.

Variable statistique qualitative ordinale. Une variable statistique qualitative ordinale est une variable dont les modalités peuvent être classées.

Variance. La variance est la moyenne des carrés des écarts des valeurs de la variable à la moyenne. Elle est notée σ_X^2 dans une population et s_X^2 dans un échantillon.

Variance corrigée. La variance corrigée est le carré de la somme des écarts à la moyenne, divisé par $(n - 1)$ où n est la taille de l'échantillon. Elle est notée \bar{s}_X^2 .

Variance intergroupe. La variance intergroupe (ou variance inter) exprime la variation entre les groupes, chacun d'entre eux étant caractérisé par sa moyenne.

Variance intragroupe. La variance intragroupe (ou variance intra) exprime la variation à l'intérieur de chaque groupe.

Variance conditionnelle. La variance d'une variable Y calculée pour une des modalités de la variable X est appelée variance

conditionnelle de Y selon la modalité x_i de X . Elle est notée $\sigma_{Y/X=x_i}^2$.

Variance marginale. Dans le cas d'une distribution conjointe, la variance d'une variable est appelée variance marginale.

Z

Zone de rejet. L'ensemble des valeurs observées de la statistique du test provoquant le rejet de l'hypothèse nulle est appelé la région critique ou zone de rejet du test statistique. Par opposition, l'ensemble des valeurs observées de la statistique du test ne permettant pas de rejeter l'hypothèse nulle est appelé zone de non-rejet.