Annexe 1 : Lexique

>> :Indique une grande valeur. Par exemple, $N >>$. Pour les estimateurs de tendance centrale (tel la moyenne, etc.), $N >>$ si $N > 30$ données brutes.

$\sum$, $\sum_{i}$, $\sum_{i=1}^{n}$: Voir sommation. Les trois symboles sont identiques; cependant le troisième indique formellement quel indice varie (*i* dans l'exemple) et combien d'éléments l'on doit sommer (*n* dans l'exemple). La première et seconde notation supposent que ces informations sont non ambiguës dans le contexte utilisé. Ex : $\sum_{i=1}^{10} i = 55$. Si les observations **X** sont {4, 8, 11, 13}, alors $\sum_{i} \mathbf{X}_i = 36$.

Dans ce deuxième exemple, l'indice *i* va de 1 à 4, puisqu'il y a quatre observations.

$\prod$, $\prod_{i}$, $\prod_{i=1}^{n}$: Voir produit. Les trois symboles sont identiques; cependant le troisième indique formellement quel indice varie (*i* dans l'exemple) et combien d'élément l'on doit multiplier (*n* dans l'exemple). La première et seconde notation supposent que ces informations sont non ambiguë dans le contexte utilisé. Ex : $\prod_{i=1}^{5} i = 120$. Si les observations **X** sont {4, 8, 11, 13}, alors $\prod_{i} \mathbf{X}_i = 4576$. Dans ce deuxième exemple, l'indice *i* va de 1 à 4, puisqu'il y a quatre observations.

! : *r*! donne la factorielle de *r*, soit $r! = r \times (r-1) \times (r-2) \times ... \times 3 \times 2 \times 1$. Par définition, 0! = 1. La factorielle est aussi parfois définie par la fonction **Γ**, tel que **Γ** (r+1) = r!

| | : Valeur absolue. | x | donne la valeur de x, dépouillée de son signe plus ou moins. Lors d'une expression telle | x – y |, il faut voir la valeur absolue comme la *distance* séparant x et y, une distance étant une valeur sans signe. Par ailleurs, deux conditions du genre « si $x > 5$ ou $x < -5$ » peut être simplifiée par « si $| x | > 5$.

$\sqrt[n]{x}$ **(racine n^ième^)** : Lorsque *n* n'est pas indiqué dans la formule, il s'agit de la racine carrée: quel est le nombre tel que ce nombre multiplié par lui-même deux fois donne *x*? Par exemple, $\sqrt{100} = 10$ car 10×10 donne 100. Si un nombre *n* est indiqué (par exemple, la racine cubique quand *n* est 3), il faut trouver le nombre tel que ce nombre multiplié par lui-même trois fois donne *x*. Par exemple, $\sqrt[3]{1000} = 10$ car $10 \times 10 \times 10$ donne 1 000.

Alphabets : Différents alphabets sont utilisés pour représenter les concepts en statistiques. Dans ces notes de cours, nous allons utiliser l'alphabet grec (voir plus loin) pour représenter des paramètres d'une population (souvent inconnus), l'alphabet cursif (tel $\mathcal{N}$, $\mathcal{B}$, $\mathcal{W}$ etc.) pour représenter des distributions, et le gras majuscule (tel **X**, $\mathbf{X_i}$, **Y**, etc) pour représenter des variables aléatoires.

Alphabet grec : Les lettres grecques sont (majuscule et minuscule) : *A*, α (alpha); *B*, β (beta); *Γ*, γ (gamma); *Δ*, δ (delta); *E*, ε (epsilon); *Z*, ζ (zêta); *H*, η (êta); *I*, ι (iota); *K*, κ (kappa); *Λ*, λ (lambda); *M*, μ (mu);

N, ν (nu); *Ξ, ξ* (xi); *O, o* (omicron); *Π, π* (pi); *P, ρ* (rho); *Σ, σ* (sigma); *T, τ* (tau); *Y, υ* (upsilon); *Φ, ϕ* (phi); *X, χ* (chi); *Ψ, ψ* (psi); et *Ω, ω* (omega).

Arrondissement et nombre de chiffres significatifs : Lorsque l'on collecte une mesure sur un individu, l'instrument de mesure utilisé n'a jamais une précision infinie. Par exemple, pour mesurer la taille d'un individu, on utilise une règle graduée en centimètre. La mesure obtenue sera donc du genre 1m 56cm ± 0.5 cm, ou encore: 1.560 ± 0.005. Autrement dit, la dernière décimale ci-haut est incertaine. Il y a donc 3 nombres qui sont sûr, qu'on appelle le nombre de chiffres significatifs. Puisque les observations ne sont sûr qu'à trois nombres significatifs (les autres décimales étant incertaines), tous les calculs fait sur ces observations (moyenne, tests, etc.) doivent être rapportés avec trois nombres significatifs aussi. Comme les ordinateurs et les calculatrices rapportent souvent plus de décimales (ils ne savent pas où débutent les nombres incertains), vous devez arrondir en conséquence. En règle générale, les instruments utilisés en psychologie ont de deux à trois nombres significatifs. Pour des besoins très pointus cependant, il est possible de construire des instruments de mesure plus précis.

CDF (*Cumulative distribution function*) : Traduit par Fonction cumulative de distribution, ou plus simplement par Fonction de distribution, représente la probabilité qu'une valeur **X** échantillonnée par hasard dans une population soit égale ou moindre que x (noté par le raccourci Pr { $\mathbf{X} \leq x$ }). Le type de fonction cumulative de distribution utilisé dépend des postulats posés sur la population. La Figure 1 illustre une fonction de distribution Normale avec paramètres $\mu = 0$ et $\sigma = 1$ (c. à. d. la Normale standardisée).

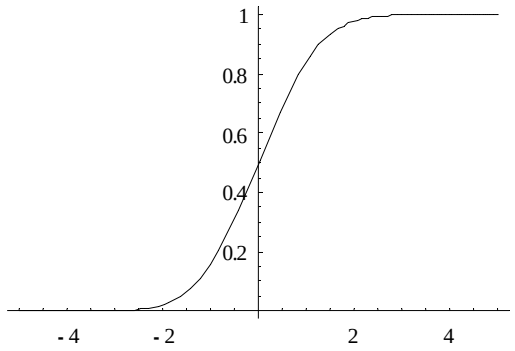

Figure 1 : La fonction cumulative de la distribution normale standardisée

Constante : Quantité qui ne change pas, qui demeure invariable. Souvent, les constantes sont représentées par les lettres *c, k, l, m* (bien que ce soit un choix arbitraire). Par exemple, ***e*** est la constante pour les logarithmes népériens, égal à 2.71828182845904523… La valeur d'une constante ne change pas à l'intérieur d'un contexte donné, tout comme la valeur de ***e*** ne change pas dans le contexte d'un logarithme.

Critère de décision: Le critère de décision est souvent notée par la lettre grecque α. Il s'agit d'une valeur choisie par l'expérimentateur (et donc constante à l'intérieur de son rapport) indiquant la probabilité que ses conclusions soient erronées (dans le sens d'une erreur α). Par exemple, si l'expérimentateur choisi un seuil α de 0.05 (valeur usuellement choisie), l'expérimentateur accepte de se tromper de conclusion dans 5 pour cent des cas. Plus α est faible, plus le critère est difficile à excéder, menant à des décisions très conservatrices. Puisque α

est choisi à priori, la valeur est souvent un nombre facile à exprimer. Par exemple, 0.05, 0.01 ou encore 0.001.

Distribution théorique :Une distribution est une fonction qui donne la probabilité d'un événement **X**. Par exemple, lorsque l'on dit que **X** est distribué comme une normale standardisée (que l'on note $\mathbf{X} \sim \mathcal{N}(0,1)$), cela implique qu'il existe une formule pour connaître la probabilité que **X** prenne la valeur x, que l'on note $\Pr\{\mathbf{X} = x\}$, et qu'on illustre avec la PDF. On a aussi souvent la fonction qui donne $\Pr\{\mathbf{X} \leq x\}$, illustrée par la CDF. Certaines distributions importantes sont la distribution Binomiale ($\mathcal{B}$) découverte par Bernoulli en 1713, la distribution Normale ($\mathcal{N}$) découverte par De Moivre en 1756, puis redécouverte par Gauss en 1800, et la distribution de Weibull ($\mathcal{W}$), découverte par Fisher et Tippett en 1928.

Donnée, donnée brute : Une mesure obtenue de la population cible. Un ensemble de données constitue l'échantillon. Souvent, pour différencier les données brutes, on va les assigner à un ensemble d'observation **X**, tel que la première mesure obtenue sera notée $\mathbf{X}_1$, la seconde $\mathbf{X}_2$, la $i^{\text{ème}}$ $\mathbf{X}_i$, et finalement, la dernière, $\mathbf{X}_n$. Ici, on suppose que n dénote le nombre total d'observations échantillonnées, et est une constante dans le contexte d'une expérience. On peut aussi voir l'échantillon comme un ensemble $\{\mathbf{X}_i\}$, où i prend les valeurs de 1 à n.

Écart type non biaisé d'un échantillon ${}_{n-1}\vec{\tilde{X}}$ ou S_{n-1}: L'écart type non biaisé d'un échantillon est donné par la racine carrée de la variance non biaisée d'un échantillon. Parce que l'écart type s'exprime dans la même unité de mesure que les données brutes, on rapporte plus souvent cette mesure que la variance.

Écart type d'un échantillon ${}_{n}\vec{\tilde{X}}$ ou S_n: L'écart type d'un échantillon est donnée par la racine carrée de la variance d'un échantillon. Parce que l'écart type s'exprime dans la même unité de mesure que les données brutes, on rapporte plus souvent cette mesure que la variance.

Échantillon : Ensemble de données brutes (observations) extrait d'une population cible. En méthodologie scientifique, l'on distingue différents types d'échantillon : L'échantillon aléatoire implique que la sélection des observations se fait entièrement par hasard, tous les éléments de la population ayant une chance égale d'apparaître dans l'échantillon. Dans l'échantillon contrôlé, les éléments sont choisis selon certains critères préétablis (par exemple, dans une étude sur des étudiants universitaires, l'échantillon sera formé en choisissant un nombre égal d'étudiants au premier, second, et troisième cycle). Deux échantillons sont appareillés lorsque la sélection des sujets se fait par paires identiques sur certains points (par exemple, deux groupes peuvent contenir le même nombre d'individus de même sexe, de même âge, etc.).

Erreur β : Une erreur dans la conclusion d'une recherche qui survient lorsque le chercheur conserve comme étant vrai son hypothèse de recherche H_0 alors qu'elle est fausse dans la population en général. Le risque que cette erreur survienne est difficilement quantifiable (voir section 13). La meilleure façon de réduire le risque de commettre une erreur β est d'augmenter la taille de

l'échantillon.

Erreur α : Une erreur dans la conclusion d'une recherche qui survient lorsque le chercheur rejette comme étant fausse son hypothèse de recherche H_0 alors qu'elle est vraie dans la population. Le risque que cette erreur survienne est quantifiable et choisi à priori par le chercheur en ajustant son critère de décision.

Erreur type : L'erreur type est une mesure de l'erreur d'estimation. Par exemple, lorsque nous calculons la moyenne d'un échantillon, nous cherchons souvent à estimer la moyenne de la population. Or, à cause d'erreur d'échantillonnage, notre moyenne estimée risque fort de ne pas être parfaitement égale à la moyenne de la population entière. Cependant, il est possible d'estimer notre erreur grâce à l'erreur type. L'erreur type devrait toujours se trouver sous forme de barres d'erreur dans un graphe rapportant des moyennes.

Hypothèse nulle H_0 : L'hypothèse nulle est toujours l'hypothèse de recherche qui affirme une absence d'effet de la ou des variables indépendantes faisant l'objet de la recherche. Les méthodes de statistique inductive ne pouvant que déceler des différences, elles peuvent soit a) rejeter l'hypothèse nulle, ou b) ne pas rejeter l'hypothèse nulle. En soi, il est impossible d'avoir une preuve empirique en faveur de l'hypothèse nulle car l'approche est basée sur un échantillonnage qui peut être non représentatif, biaisé ou encore trop petit.

Intégrale, $\int$: Opérateur mathématique permettant de calculer l'aire sous une courbe. Cette opération est très similaire à une sommation dans laquelle les rectangles à additionner sont très étroits, et souvent une sommation peut être utilisée pour trouver de façon approximative la valeur numérique d'une intégrale.

Logarithme : La fonction logarithme ***log(x)*** donne le nombre y tel que 10^y donne x. Par exemple, ***log(1 000)*** = 3 car 10^3 donne $10 \times 10 \times 10$, soit 1 000. La fonction ***ln(x)*** est semblable, exceptée qu'elle utilise la base ***e*** =2.71828 plutôt que 10.

Médiane $\breve{X}$: La médiane est une valeur qui divise un ensemble d'observations en deux sous-ensembles de taille égale tel que tous les éléments du premier sous-groupe sont inférieurs à la médiane, et tous les éléments du second sous-groupe, supérieurs. Pour calculer la médiane sur un échantillon, il faut trier les observations en ordre croissant, puis a) si n est impair, choisir la donnée qui se trouve au centre de la série, ou b) si n est pair, choisir la moyenne des deux observations les plus centrales. Par exemple, étant donné l'échantillon (déjà trié) **X** = { 2, 4, 6, 7, 8, 9, 11 }, la médiane est 7. Pour calculer la médiane à partir d'un graphique cumulatif des fréquences, trouver la valeur x tel que la fréquence observée soit de 50%. Par exemple, sur la Figure 1 ci-haut, la valeur médiane est 0.

Mode $\dot{X}$: Le mode correspond au score que l'on retrouve le plus souvent dans un échantillon. Les échantillons peuvent être unimodals (par ex. **X** = {1, 3, 3, 3, 4, 4, 6 }), plurimodals (par ex. **X** = {1, 3, 3, 3, 4, 4, 4, 6 }) ou encore amodals (par ex. **X** = {1, 3, 4, 5, 6 }). Si les données sont présentées sous forme de graphique en histogramme, le mode correspond à l'histogramme le plus élevé.

Moyenne arithmétique $\overline{X}$: La moyenne arithmétique (le type de moyenne le plus

souvent utilisé). La moyenne se calcule en faisant la somme des observations, puis en divisant par le nombre d'observées, noté $\frac{1}{n}\sum_i \mathbf{X}_i$. Par exemple, si **X** = { 2, 4, 6, 7, 8, 11, 18 }, $\overline{X}$ = 8. À partir d'un graphique en histogramme, la moyenne est le point sur l'abscisse qui tient la distribution en équilibre (c. à d. le centre de gravité de la distribution).

Moyenne géométrique $\overset{\circ}{X}$: La moyenne géométrique est la $n^{\text{ième}}$ racine du produit des observations, notée $\sqrt[n]{\prod_i \mathbf{X}_i}$. Par exemple, si **X** = { 2, 4, 6, 7, 8, 11, 18 }, $\overset{\circ}{X}$ = 6.57. La moyenne géométrique est utilisée dans certains contextes de recherche précis, par exemple lorsque les données possèdent beaucoup de valeurs à une extrémité seulement. L'idée sous-jacente à ce type de moyenne vient du fait qu'on peut rendre les données extrêmes moins extrêmes en considérant le log des données. Soit $\mathbf{Y}_i = ln\ \mathbf{X}_i$. Si l'on calcule $\overline{Y} = \frac{1}{n}\sum_i \mathbf{Y}_i$, l'on obtient :

$$\begin{aligned}\overline{Y} &= \frac{1}{n}\sum_i \ln \mathbf{X}_i \\ &= \frac{1}{n}\ln \prod_i \mathbf{X}_i \\ &= \ln \sqrt[n]{\prod_i \mathbf{X}_i} \\ &= \ln \overset{\circ}{X}\end{aligned}$$

d'où il s'ensuit que $\overset{\circ}{X} = e^{\overline{Y}}$.

Notez que la moyenne géométrique ne fait pas de sens pour des nombres négatifs. Par ailleurs, si vous avez des zéros, vous devez les remplacer par la plus petite valeur non nulle que votre instrument de mesure peut retourner.

Moyenne harmonique $\tilde{\tilde{X}}$: La moyenne harmonique correspond à l'inverse multiplicatif de la moyenne des inverses des données brutes, notée $n \Big/ \sum_i \frac{1}{\mathbf{X}_i}$. Par exemple, si **X** = { 2, 4, 6, 7, 8, 11, 18 }, $\tilde{\tilde{X}}$ = 5.26. L'idée sous-jacente à ce type de moyenne vient du fait qu'on peut considérer des ratios (des taux de changement) plus efficacement en considérant l'inverse multiplicatif des données. Soit $\mathbf{Z}_i = 1 / \mathbf{X}_i$. Si l'on calcule $\overline{Z} = \frac{1}{n}\sum_i \mathbf{Z}_i$, l'on obtient :

$$\overline{Z} = \frac{1}{n}\sum_i 1/\mathbf{X}_i$$

d'où il s'ensuit que $\tilde{\tilde{X}} = 1/\overline{Z} = \frac{n}{\sum_i 1 \Big/ \mathbf{X}_i}$

Notation scientifique : Façon d'écrire des nombres lorsqu'ils sont très grands ou très petits, utilisé entre autre par les calculatrices. Soit par exemple, le grand nombre qui suit: 345 000 000. On peut de façon équivalente, écrire 3.45×10^8, c'est à dire 3.45 multiplié par 100 millions. Bien que les deux nombres soient égaux, le second est plus court à écrire et utilise la notation scientifique. Une façon rapide de lire cette notation est de simplement prendre l'exposant (8 ici) et de déplacer le point décimal de 8 positions vers la gauche, en rajoutant des zéros si nécessaire. Si l'exposant est négatif, on déplace la virgule vers la droite. Par exemple, le petit nombre 0.000 000 034 5 s'écrit en notation scientifique 3.45×10^{-8}. On appelle parfois la premier partie (3.45 ici) la mantisse pour la distinguer de l'exposant. Sur une calculatrice, seul la mantisse et l'exposant sont affichés, séparés souvent par la lettre "E", ce qui

donne 3.45 E –8.

Paramètre d'une population : Le paramètre est une valeur caractérisant une population. Il représente souvent une quantité inconnue que l'on tente d'estimer au moyen de méthodes statistiques. Par exemple, la taille moyenne μ d'une population d'individus est inconnue, mais on l'estime à l'aide de la moyenne arithmétique $\overline{X}$ d'un échantillon. Les paramètres d'une population sont souvent dénotés par des lettres grecques (par exemple, μ et σ).

PDF (*Probability density function) :* Traduit par Fonction de masse, représente la probabilité qu'une valeur **X** échantillonnée par hasard d'une population soit égale à x (noté par le raccourci Pr { **X** = x }). Le type de Fonction de masse à utiliser dépend des postulats posés sur la population. La Figure 2 illustre une fonction de distribution Normale avec paramètres $\mu = 0$ et $\sigma = 1$ (c. à. d. la Normale standardisée).

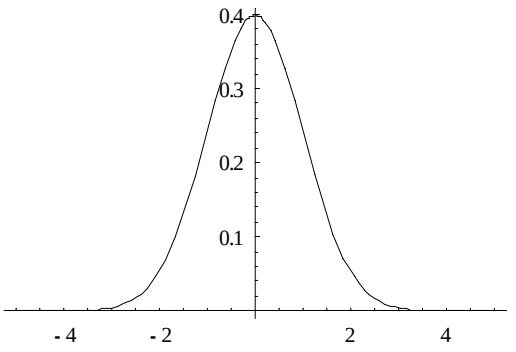

Figure 2 : Fonction de masse de la distribution normale standardisée

Population : Ensemble de toutes les observations possibles concernant l'objet d'étude. À cause des limites financières et temporelles, il est généralement impossible de faire l'étude exhaustive d'une population complète. Par exemple, la population ciblée par une recherche pourrait porter sur l'ensemble des enfants de cinq ans habitant l'Amérique.

Produit : Multiplication d'une suite de termes, et noté par le symbole Pi majuscule, $\prod$.

Sommation : Addition d'une suite de termes, généralement des éléments tirés d'un ensemble de valeur. La sommation est très utilisée en statistique, et pour faciliter sa notation, le symbole $\sum$ est utilisé.

Statistique d'un échantillon : Valeur extraite d'un échantillon et sensé être représentatif de la population ou pouvant être utilisée dans un test statistique. Par exemple, la taille moyenne d'un échantillon d'enfants de cinq ans habitant l'Amérique est une statistique.

Statistique descriptive : Branche des statistiques servant à décrire un échantillon, soit en utilisant des graphiques (tel le graphique des histogrammes) ou en calculant des valeurs représentant les données de façon condensée (*summary values*).

Statistique inductive : Branche des statistiques permettant de tirer des conclusions sur une population à partir de l'étude d'un échantillon représentatif (aussi appelé statistique inférentielle).

Statistiques (pl.) : Branche des mathématiques qui étudie les propriétés d'échantillons extraits d'une population plus large, en utilisant des postulats de base décrivant la population, et la taille de l'échantillon. Les statistiques s'opposent aux probabilités, qui étudient les propriétés d'une population entière,

étant donné des postulats de base. Souvent, les probabilités sont complémentaires des statistiques car les postulats de base sont parfois vagues et leurs conséquences sont explorées par les probabilistes (en terme de CDF et PDF).

Test statistique : Méthode de la statistique inductive pour faire des inférences sur la valeur d'un paramètre d'une ou plusieurs populations à partir d'une ou plusieurs statistiques obtenues d'un échantillon. Par exemple, un test peut être utilisé pour vérifier si la taille moyenne des nord-américains est supérieure à la taille moyenne des africains, étant donné que nous avons accès à un échantillon de 500 africains et 500 nord américains.

Variable aléatoire : Attribut ou caractéristique d'une population qui peut faire l'objet d'une mesure. Les mesures, ou observations (données brutes), sont collectées dans un ensemble qui est représenté par une lettre (toujours gras et majuscule). Souvent, on utilise **X** ou **Y**, quoique ce choix soit arbitraire. Ce type de variable est dit aléatoire car les valeurs exactes ne sont jamais connues à l'avance. Par exemple, soit **X** l'ensemble contenant la taille des étudiants de cette classe.

Variable dépendante : La mesure obtenue de la population qui est le sujet des hypothèses du chercheur.

Variable indépendante : Le ou les facteurs que le chercheur manipule (méthode expérimentale) ou met en relation (méthode corrélationnelle) pour évaluer les effets sur la variable dépendante. Par exemple, quels sont les effets de la qualité de la nutrition et de la quantité d'exercice sur la taille atteint à l'âge adulte. Ici, la taille à l'âge adulte est la variable dépendante. Si l'étude est effectuée sur des humains, il s'agit sans doute (pour des raisons d'éthique) d'une étude corrélationnelle.

Variance non biaisée d'un échantillon $_{n-1}\vec{\vec{X}}^2$ ou S_{n-1}^2: La variance d'un échantillon est donnée par la moyenne des écarts à la moyenne mis au carré, et corrigée pour un biais (dû à la petitesse de l'échantillon) par le facteur $\frac{n}{n-1}$. On remarque que si l'échantillon est infini, la correction $\underset{n\to\infty}{Lim}\frac{n}{n-1}\to 1$ s'amenuise. La formule finale est $\frac{1}{n-1}\sum_i(\mathbf{X}_i-\overline{\mathbf{X}})^2$. La variance étant un carré, elle s'exprime par l'unité de mesure des données brutes au carré. Par exemple, si les données sont des mètres, la variance s'exprime en mètres carrés. Pour cette raison, l'on va préférer rapporter l'écart type, donné par la racine carrée de la variance. Par exemple, si **X** = { 2, 4, 6, 7, 8, 11, 18}, alors $_{n-1}\vec{\vec{X}}^2$ = 27.7.

Variance d'un échantillon $_{n}\vec{\vec{X}}^2$ ou S_n^2: La variance d'un échantillon est donnée par la moyenne des écarts à la moyenne mis au carré, noté $\frac{1}{n}\sum_i(\mathbf{X}_i-\overline{\mathbf{X}})^2$. La variance étant un carré, elle s'exprime par l'unité de mesure des données brutes au carré. Par exemple, si les données sont des mètres, la variance s'exprime en mètres carrés. Par exemple, si **X** = { 2, 4, 6, 7, 8, 11, 18}, alors $_{n}\vec{\vec{X}}^2$ = 23.7. Remarquez que $\frac{n}{n-1}\ {}_{n}\vec{\vec{X}}^2 = {}_{n-1}\vec{\vec{X}}^2$.