BALISE π[1] : STATISTIQUES

Les statistiques permettent le traitement des données en vue de leur analyse ; la statistique qui conduit à l'élaboration des statistiques dont la représentation la plus communément partagée est celle du graphique. Cette étape, de traitement et d'analyse des données recueillies par les outils d'observation et d'enquête que nous étudierons plus loin, n'est que le préalable à l'interprétation de ces données. C'est là que la méthode statistique cède le pas à la méthodologie du discours. Cette « grosse[2] » balise sera subdivisée en chapitres dont l'étude sera étalée en deux étapes.

Contenu de cette macro-balise :

1. Généralités statistiques.
2. Stat… avec ou sans S.
3. Usage statistique.
4. Usage terminologique.
5. De la variable en passant par l'échelle.
6. De l'échantillonnage.
7. Tableau de répartition.
8. Représentations graphiques.
9. Les indices de résumé statistique.
10. Corrélations.
11. Représentation graphique, communication et autres manœuvres…
12. Notion de loi statistique.
13. Tests paramétriques ou tests d'hypothèse.
14. Variables qualitatives et Chi carré.
15. Estimation ou intervalles de confiance.

[1] Souvenir du collège, cette lettre *pi* est aussi un nombre qui intervient dans le calcul de la circonférence, ou périmètre comme de l'aire, ou surface d'un cercle. Rappel, encore : le cercle peut circonscrire un triangle.

[2] Le volume de cette balise n'est pas un choix éditorial de l'auteur, mais à mettre en relation avec les contraintes du dossier pédagogique de cette formation qui alloue aux statistiques quelques 30 heures et d'autre part, l'usage et la préséance encore trop souvent accordés au modèle expérimental et aux chiffres qui le soutiennent.

Chapitre 1 : Généralités.

Le type de traitement des données est en lien étroit, et donc congruent, avec la méthode et les objectifs de la recherche. Les méthodes expérimentale et différentielle privilégient une approche quantitative des données, mais qui peut être complétée par une approche qualitative. Les deux approches peuvent avoir leur utilité dans les méthodes socio-anthropologiques et historique. En méthode clinique, c'est l'interprétation clinique qui est au centre de la démarche, mais il est toujours possible d'évaluer quantitativement l'un ou l'autre élément du discours.

Les outils de recueil « typiques » d'une approche quantitative sont le questionnaire et la grille d'observation. L'analyse quantitative exige alors la réduction du contenu en énoncés clairs, la rédaction d'indicateurs et leur codage. Ces opérations se font avant l'enquête, *a priori* ; le traitement et l'analyse sont prédéterminés par le dispositif de recherche.

Dans le cas d'entretiens semi-directifs[3], il est aussi possible de quantifier comme l'analyse de contenu le suggère. L'analyse du discours calcule la fréquence d'apparition des unités de sens. Toutefois, le comptage s'opère sur la similitude de formes, le travail sur le sens[4] et l'interprétation ne se font que par le chercheur.

La statistique est un ensemble de techniques et de méthodes permettant d'analyser des observations transformées en données, données concernant un fait. Le traitement statistique, basé sur l'arithmétique et la géométrie, peut s'intéresser à la description d'une ou de plusieurs variables. Chaque variable peut être étudiée séparément. L'exemple-type est d'analyser la répartition de la population en fonction des valeurs de chaque caractère pris successivement. La statistique permet également, et là est l'optimum, de décrire les liaisons entre plusieurs caractères et/ou variables. Le traitement statistique des données s'appuie toujours sur la statistique descriptive et, suivant le type d'enquête, sur l'inférence statistique.

- *L'analyse descriptive*, première, a pour but de décrire un phénomène en résumant ses caractéristiques quantitatives en quelques nombres.
- *L'inférence statistique*, seconde, propose de déduire les caractéristiques d'une population parente ou théorique à partir de l'étude d'un échantillon. Elle se propose aussi de mettre en évidence le type de relations entretenues par les phénomènes observées, les caractères et/ou les dimensions d'une ou de plusieurs variables. Cette étape est incontournable dans l'expérimentation d'une hypothèse de recherche.

[3] Nous reviendrons plus loin sur cette notion. Sachez toutefois qu'un entretien fermé ou directif se traite comme un questionnaire.

[4] Cette technique s'appelle l'analyse de contenu, opère par regroupement en catégories. Nous l'étudierons l'an prochain.

Chapitre 2 : Stat… avec ou sans S.

Avant d'aborder ces différents outils proposés par la statistique, il est nécessaire de distinguer son singulier du pluriel. Il est, ici et maintenant, utile de rappeler l'origine étymologique du mot « statistique ». Il remonte au latin classique *status*, état/Etat, qui, par une série d'évolutions successives, aboutit au français *statistique*, attesté pour la première fois en 1771. C'est vers la même époque que *statistik* apparaît en allemand, alors que les anglophones utilisent l'expression *political arithmetic* jusqu'en 1798, date à laquelle le mot *statistics* fait son entrée dans cette langue. A l'origine, cette discipline concerne donc les affaires de l'Etat.

Actuellement, on distingue généralement les statistiques, au pluriel, de la statistique, son singulier. Les statistiques peuvent être définies comme l'étude méthodique des faits sociaux qui définissent un Etat, par des procédés numériques comme dénombrements, inventaires, recensements, etc. . Le sens commun ne retient encore que la représentation graphique de ces exercices. Le second sens, le singulier, n'apparaît que vers 1830. La statistique, abordée dans ce cours, se définit comme un ensemble de techniques d'interprétation mathématique appliquées à des phénomènes pour lesquels une étude exhaustive de tous les facteurs est impossible à cause de leur grand nombre ou de leur complexité. La statistique peut donc être considérée comme le processus qui conduit à un produit, les statistiques.

De tous temps, les chefs d'Etat ont souhaité déterminer la puissance des nations qu'ils dirigeaient à l'aide de recensements partiels ou complets de leur : population, territoire, production, etc. . Déjà en 3000 av. J.-C., en Mésopotamie, on dresse inventaire. Mais surtout, au début de notre ère, a lieu un dénombrement des richesses de l'Empire Romain, rendu célèbre par sa mention dans l'Evangile de Luc. Au XVII^{ème} Siècle, pour éviter les recensements, lourds et onéreux, William Petty met au point une méthode de comptage de la population de Londres sur base des proportions moyennes entre les maisons, les feux[5] par maison et la composition des familles.

Au XIX^{ème} Siècle, les recensements proprement dits reprennent de l'importance et, en 1853, a lieu à Bruxelles le 1er Congrès International de Statistique, sous l'impulsion d'Adolphe Quételet[6]. L'objectif de ce congrès est d'uniformiser, de standardiser les techniques de compilation des statistiques nationales, en vue de faciliter les comparaisons.

Au début du XX^{ème} Siècle, un débat oppose les partisans des recensements, c'est-à-dire des comptages réalisés sur l'ensemble de la population et les adeptes du sondage, réalisé sur un échantillon de cette même population. Les recensements ne sont pas toujours possibles, ni

[5] Feux - foyers - ménages ou familles ! Ces notions se retrouvent en « Sécurité Sociale ».
[6] 1796-1874, astronome et mathématicien belge, il est considéré comme un des fondateurs de la science statistique.

souhaitables. Dans certains cas, ils peuvent être trop chers. Ils peuvent aussi contenir des erreurs. Parfois, ils sont carrément aberrants[7].

Pour pallier ces inconvénients, on a recours au sondage statistique, qui consiste à déduire les propriétés de toute une population à partir de l'analyse d'un échantillon. Il est capital que l'échantillon soit choisi et analysé de manière adéquate. En particulier, il faut que l'échantillon soit représentatif de la population. S'il ne l'est pas, il est dit biaisé[8].

Nous retiendrons ici une des deux notions fondamentales de la démarche statistique : la population et son échantillon. La seconde notion constitutive est celle de variable.

[7] Mesurer la solidité moyenne d'un type de voiture en lançant toutes les voitures de ce type contre un mur serait commercialement inacceptable

[8] Au début du XX^{e} siècle, les journaux américains réalisent des « votes de paille » en demandant l'avis par écrit à plusieurs millions de personnes quelques semaines avant les élections présidentielles. En 1936, le *Literary Digest* prédit, à l'aide d'un échantillon de 2.400.000 électeurs, la victoire du candidat républicain. G. Gallup, grâce à un sondage sur 4000 personnes **judicieusement** choisies, prévoit, quant à lui, la victoire du démocrate F.D. Roosevelt. La victoire de ce dernier sonne le glas des votes de paille dont les échantillons sont souvent biaisés (les cartes du *Literary Digest* avaient été envoyées aux abonnés du téléphone et aux propriétaires de voitures, cet électorat aisé était plus favorable aux républicains).
En réalité, et particulièrement en sciences sociales, tout échantillon est biaisé. Certains biais peuvent être contributifs à la recherche.

Chapitre 3 : Usage statistique.

Dans le cadre de l'utilisation d'une méthode, d'un outil, d'une science, il est vivement recommandé d'utiliser la terminologie retenue et partagée au sein de ce contexte. D'ailleurs, l'utilisation de ces termes relève de la conceptualisation[9]. « *The right word in the right use* ».

Si vous utilisez la statistique dans le cadre de votre EI, il devient incontournable d'utiliser la terminologie statistique. La compréhension du sens de ce terme conditionne, outre la formulation du discours statistique, mais aussi, le choix des outils que propose la méthode statistique. Une précaution d'usage encore suit.

La précision des chiffres et la rigueur de représentation des statistiques leurs confèrent une apparente objectivité. Son origine comme sa crédibilité résident d'abord dans l'arithmétique, considérée comme une science exacte ; où « 1 + 1 = 2 ». D'autre part, expliquant en partie la préséance mentionnée plus haut, son usage intempestif dans la médecine contemporaine a rendu le modèle hypothético-déductif ou expérimental comme principe fondateur de cette science, pourtant humaine. Par contagion ou au moins par proximité, le modèle médical aurait-il déteint sur nos pratiques de soignants comme d'encadrement ? Le danger ne demeure pas dans l'usage de la statistique et/ou de l'expérimentation mais lorsqu'elles sont considérées comme seule voie de légitimation d'un savoir ; délaissant ainsi, voire dénigrant par-là, les autres méthodes et pratiques.

La technologie ne peut se substituer à la méthodologie. Le moyen n'est rien sans la fin. Comme tout outil, certaines limites réclament une dose de prudence dans son utilisation. Ce discernement se retrouvera encore lorsque nous approcherons la communication didactique et la présentation visuelle de la représentation graphique de statistiques. Enfin, et non des moindres en ce qui nous concerne, dernières mesures (c'est le cas de le dire !) :

- les faits qui vous intéressent ne sont pas toujours mesurables quantitativement ;
- la statistique a un pouvoir d'élucidation limité aux hypothèses[10] sur lesquels il repose, mais ne disposent pas, en elles-mêmes, d'un pouvoir explicatif ;
- la statistique et les statistiques n'interprètent pas, c'est le chercheur qui donne un sens à ces chiffres en lien avec le modèle conceptuel développé en amont et en fonction des méthodes d'analyse statistique qu'il choisit.

L'outil n'est rien sans l'artisan.

[9] Autrement dit, un terme possède généralement plusieurs sens, il est polysémique. La signification est attribuée, entre autre, en fonction du contexte. Son usage est directement conditionné par ce sens.
[10] Retournez lire la définition de l'hypothèse.

Chap. 4 : Usage terminologique.

La « bonne » utilisation des mots que cette terminologie propose, conditionne la pertinence du choix des méthodes statistiques étudiées : calcul ou pas calcul, tel test ou l'autre, etc. Nous le savons ; deux éléments sont à la base du raisonnement statistique : la population et la variable, notions considérées comme constituantes. Le premier renvoie à la technique de l'échantillonnage et à son « adret » : l'estimation. Elle est une forme de retour sur la population au départ de l'échantillon dans la perspective de cette méthode : la généralisation et l'énoncé d'une loi ! Quant à elle, la notion de variable doit s'entendre et s'utiliser ici dans le cadre des statistiques : une variable statistique ; qui ne faut pas confondre avec une variable d'hypothèse !

Au sens de la méthode expérimentale, l'hypothèse[11] se compose d'au moins deux variables reliées entre elles par un lien de causalité, supposé et à démontrer par l'expérimentation. Celles-ci sont bien souvent déterminées comme dépendante pour la première[12], et d'indépendante pour la seconde.

Chaque variable de l'hypothèse est déclinée en un *certain*[13] nombre d'indicateurs, eux-mêmes subdivisée en un *certain* nombre d'indices puis d'items. Ces indicateurs, indices et items peuvent devenir des variables statistiques s'ils autorisent une mesure. L'opérationnalisation permet de passer d'une variable d'hypothèse à une (plusieurs) variable(s) statistique(s).

Deux types de mesures, et donc de variables statistiques, sont envisageables : la variable peut être qualitative ou[14] quantitative. Si le caractère est qualitatif, vous serez contraint de le quantifier (*sic*). Les variables quantitatives peuvent encore être qualifiées de discrètes ou de continues. La variable est dite discrète si elle ne prend que des valeurs isolées. La variable est dite continue si elle peut prendre toutes les valeurs d'un intervalle. La variable est ainsi déterminée par deux adjectifs, allant conditionner le choix et l'utilisation des méthodes statistiques comme leur représentation graphique. Nous approcherons ces distinctions dès le chapitre suivant : les échelles.

Notons toutefois qu'outre le lien de causalité, considéré comme raison ultime de la méthode expérimentale, il est possible de travailler sur deux variables statistiques simultanément en utilisant la corrélation et sa représentation graphique : le nuage de points. Cette méthode statistique s'utilise dans une approche moins expérimentale, moins causaliste du lien. En effet, une corrélation entre deux variables peut être interprétée comme une co-variation entre composantes d'un même système ; elles évoluent conjointement. Un chapitre y est consacré.

---

[11] L'hypothèse a un sens d'usage philosophique que l'on retrouve dans une famille : thèse, hypothèse, antithèse, synthèse… devenu par débordement une notion polysémique.
[12] C'est la variable d'effet.
[13] cf. schéma d'arborescence in « Balise *ksi* ».
[14] Ce *ou* est exclusif.

Chapitre 5 : De la variable en passant par l'échelle.

Le traitement quantitatif d'une variable nécessite son opérationnalisation en différents caractères mesurables indiquant le niveau de variation d'une information. L'échelle permet de mesurer, de quantifier le niveau de variation du caractère d'une variable.

Trois types d'échelles de mesure peuvent être distingués : les échelles nominales, les échelles ordinales et les échelles d'intervalle. Chacune d'elles a des propriétés différentes ne permettant pas le même traitement statistique. En effet, leurs caractéristiques sont directement dépendantes du type de variable considérée : qualitative versus quantitative, discrète ou continue.

1. <u>L'échelle nominale</u> :

   Pour construire ce genre d'échelles, il convient de répartir les observations ou les caractères d'une variable dans un certain nombre de classes. Ces classes ne sont pas présentées dans un ordre particulier, n'ont pas de lien spécifique entre elles.

   Exemples : catégories socioprofessionnelles, sexe, type de voiture, etc.

   Les propriétés d'une échelle nominale sont :

   - chaque observation doit rentrer dans une classe et une seule,
   - deux observations appartenant à la même classe sont considérées comme équivalentes,
   - chaque observation doit pouvoir être classée dans une des catégories de l'échelle,
   - chacun des éléments peut être désigné par un nombre. Ce nombre n'est qu'un code et non une échelle de valeurs ; il indique seulement des réponses différentes,

     Exemple : homme = 1 / femme = 2
   - le nombre de données appartenant à chaque classe constitue l'effectif de cette classe,
   - la distribution des effectifs est le tableau qui représente le nombre d'observations par classe.

2. <u>L'échelle ordinale</u> :

   Elle établit une relation d'ordre linéaire entre les observations ou les caractères de la variable étudiée. Le classement peut se faire par ordre croissant ou décroissant.

   Exemple : jamais - rarement - parfois – souvent - très souvent - toujours

   Les propriétés d'une échelle ordinale sont :

   - les effectifs peuvent être cumulés de façon ordonnée,

     Exemple : {jamais – rarement} – {parfois – souvent} – {très souvent – toujours}
   - il est possible de comparer n'importe quel caractère ou sujet par rapport à un autre,
   - les caractères n'entretiennent qu'une relation de transitivité,

     A > B et B > C alors A > C

- les nombres ne servent que de code à des échelles de mots ; aucun calcul arithmétique ou statistique n'est alors envisageable.

3. Les échelles numériques continues ou discontinues :
   - Dans une échelle numérique discontinue, le caractère observé ne peut prendre que des valeurs numériques isolées, il est quantitatif discontinu[15],

     Exemple : combien d'enfants avez-vous ?
   - Dans une échelle numérique continue, aussi appelée échelle d'intervalle, le caractère observé peut prendre toutes les valeurs d'un intervalle défini. La mesure peut exister[16] ou être construite. L'intervalle doit être identique[17] tout le long de l'échelle.

     Exemple : l'intervalle entre 56 et 57 kg est identique qu'entre 72 et 73 kg.

[15] dis-continue comme dis... !
[16] le poids en kg.
[17] Dans sa mesure, il n'en sera pas nécessairement de même dans sa représentation graphique.

Chapitre 6 : De l'échantillonnage.

Réputé fondateur de l'enquête, l'échantillon reste une des difficultés majeures de l'approche expérimentale, comme de la recherche en général. Une négligence à ce niveau a des effets exponentiels au long du processus que l'échantillonnage initialise. En réalité, plusieurs échantillonnages opéreront dans une démarche.

Des termes sont, et c'est bien normal, communs tant à la population qu'à l'échantillon puisque ici, l'œuf vient de la poule ou plutôt la pomme du pommier ! L'instrument de mesure qu'est une variable statistique s'appliquera sur un *certain* nombre d'individus.

EFFECTIF	indique le nombre de fois qu'une valeur associée à un caractère a été observée (n^i).
EFFECTIF CUMULE	correspond à la somme des effectifs (= n).
FREQUENCE	est le rapport de l'effectif d'une valeur sur le nombre n de sujets de l'échantillon.
FREQUENCE CUMULEE	correspond à la somme de certaines fréquences (= 1)

L'usage de l'un comme de l'autre débouche sur un tableau de répartition permettant ensuite la présentation d'une courbe de distribution[18]. De fait, l'effectif de la population (N) conditionne la taille, l'effectif (n) de l'échantillon. La technique de constitution de l'échantillon est appelée « échantillonnage ».

Ces outils décrits, toujours de manière incomplète puisqu'à construire, ne sont pas spécifiques à la recherche, sont utilisés dans toute procédure d'enquête. L'échantillonnage est le premier instrument de cette boîte à outils, il se retrouve fréquemment premier en ordre d'utilisation sans lui ôter son caractère fondamental. En effet, cette opération, d'apparence banale, et trop souvent malmenée, permet souvent de retrouver la présence ou l'absence des critères de rigueur d'une recherche. L'échantillon est comme la partie la plus visible, lisible, de l'iceberg de l'EI, surtout lorsque la méthode se veut ou se prétend quantitative. Dans la même caisse, d'autres balises suivront :

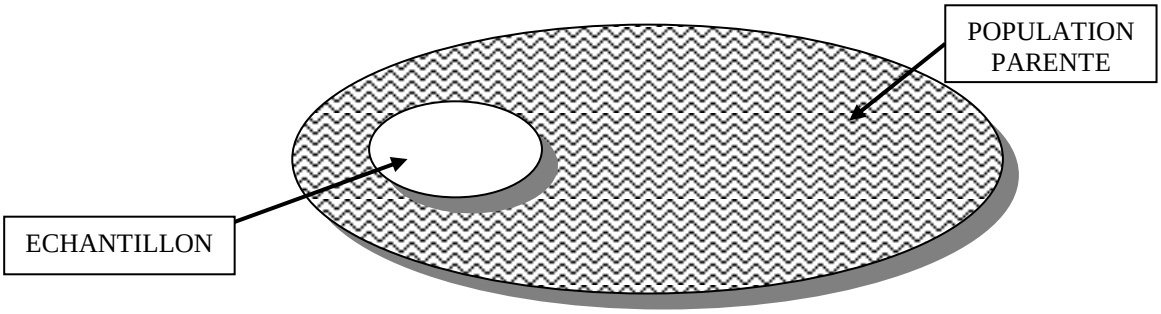
- l'usage des statistiques comme outil d'analyse et de représentation,
- le questionnaire, l'observation et l'interview comme techniques de collecte.

La question du choix des techniques et des outils se fait et s'argumente en fonction de leurs pertinences quant à l'objet étudié. Chaque technique a sa spécificité et le niveau d'information recueilli n'est pas le même suivant l'outil choisi. Si le choix est une étape essentielle, leur construction devra garantir la fiabilité et la validité du recueil, de même que sa pertinence à l'étude. De même, le choix de la population à étudier et l'identification de l'échantillon de sujets sur lequel portera l'observation sont déterminants dans la qualité d'une étude.

[18] Répartition et distribution sont des synonymes en statistique.

Echantillonnage :

Le terme échantillonnage est utilisé pour indiquer que le recueil de données se fait sur une partie de la population totale, ainsi manipulée. La population totale, théorique ou parente, est constituée de l'ensemble des personnes dont la recherche prétend rendre compte. Comme il ne vous est guère possible d'expérimenter sur tous les individus de la population, vous pouvez vous adresser à une partie d'entre eux : l'échantillon. L'échantillon est prélevé par une technique d'échantillonnage.



Pour construire un échantillon, il faut :

- choisir une population totale, en fonction du thème et de l'objectif et en déterminer les caractéristiques[19]. Elle sera théorique si vous ne pouvez la dénombrer avec précision ; elle sera parente dans le cas contraire ;
- définir une partie de cette population, de sorte que les données recueillies sur celle-ci permettent une estimation correcte de celles de la population parente. Ces marques et attributs deviennent alors des critères d'inclusion ou d'exclusion.

Il existe plusieurs façons de procéder pour construire un échantillon. L'échantillon peut être représentatif ou significatif, de même qu'il peut être exhaustif, sans remise ou non. La méthode peut également solliciter, mobiliser, manipuler deux, et plus, échantillons issus de la même population. Ces échantillons seront alors ou appareillés ou indépendants.

Echantillon représentatif ou significatif :

Deux approches sont ainsi envisageables. Un échantillon est dit *représentatif* lorsque les éléments qui constituent la population totale ont tous la même chance d'être choisis et donc de faire partie de l'échantillon. A « l'inverse », un échantillon est dit *significatif* lorsque le choix d'un sujet est argumenté en fonction de la pertinence de ce qui le caractérise ou le spécifie par rapport à l'objet de la recherche. Il est donc question d'approches.

[19] l'identité et ses signes ou attributs.

- *L'approche probabiliste* :

  Cette approche se réfère aux règles statistiques de la loi du hasard, dont nous reparlerons, et permet de construire un échantillon représentatif.

  - L'échantillon aléatoire :

    Le choix des individus se fait par tirage au sort ; il faut donc que tous les individus aient la même probabilité de faire partie de l'échantillon.

  - L'échantillon aléatoire sans contrôle des catégories :

    Il faut tout d'abord lister les personnes appartenant à la population parente, c'est la base de sondage. Le fait de tirer au hasard 100 personnes sur une base de sondage de 1000 individus permet d'avoir un échantillon représentatif. Les caractéristiques de la population parente doivent se retrouver dans cet échantillon. Certains outils existent afin de constituer ce type d'échantillon : table de chiffres aléatoires[20], algorithme informatique, par exemple.

  - L'échantillon aléatoire avec contrôle des catégories :

    Il s'agit de faire de même tout en vérifiant certaines caractéristiques fortes de la population. Celles-ci sont hypothétiquement estimées comme pertinentes en regard de la problématique étudiée. Il est dès lors requis de lister et d'argumenter le choix de ces caractéristiques[21] de manière à pouvoir les retrouver dans l'échantillon. Le tirage au sort se fait ainsi en maintenant la ou les répartition(s) exprimée(s) en pourcentage.

  - L'échantillon par grappes :

    Cette méthode d'échantillonnage ne se fait pas sur l'unité individuelle, mais sur des ensembles, des grappes d'unités voisines. Toutefois, cette technique présente l'éventualité de révéler des phénomènes minoritaires ou localisés. Pour augmenter la fiabilité de l'enquête, il est préférable d'augmenter la taille de l'échantillon.

    Par exemple : un chercheur s'intéresse à un problème de santé lié à la scolarité primaire. Il y a 150 écoles, en tire 5 au hasard et interroge tous les enfants. Il peut aussi tirer au hasard 40 écoles et interroger 15 enfants dans chacune d'elles.

  - L'échantillon stratifié :

    Il s'agit de constituer l'échantillon en prenant des proportions d'individus différentes suivant des catégories choisies : les strates. La population est divisée en strates, construites en fonction de leur pertinence. Elles regroupent les sujets sur une caractéristique commune. Pour chacune de ces strates, l'enquêteur prélève un échantillon représentatif. L'échantillon total n'est plus représentatif ; par contre, la dispersion de la caractéristique est moins grande que dans la population.

[20] Voir en fin de balise, son chapitre 16.
[21] Exemple : répartition en âge, en sexe, catégorie…

- *L'approche empirique* :

  Constructible selon trois méthodes, cette approche est contestée au niveau scientifique.

  - La méthode des quotas :

    Elle considère que certaines variables caractérisant un individu sont liées entre elles. Il faut donc choisir quelques variables identifiant de manière significative la population parente et relever leurs fréquences. La sélection se fait sur base de ces quotas préétablis. La représentativité, et donc le degré de généralisation des résultats, est proportionnelle à la pertinence des variables sélectionnées.
  - La méthode des itinéraires :

    Technique principalement utilisée par les sondeurs.
  - L'échantillon sur place :

    La constitution de l'échantillon est liée au fait de la présence des sujets dans un lieu, à un moment donné. Le lieu est élu en fonction du thème étudié. L'échantillon n'est guère représentatif, il est plutôt spatial et temporel. C'est le mode mineur de l'échantillon, il ne peut prétendre à la représentativité ; son degré de significacité est aléatoire.

Taille de l'échantillon :

La taille de l'échantillon n'est nécessairement conditionnée par la taille de la population-mère. Cet effectif est également déterminé par la technique d'investigation, la technique d'échantillonnage et la qualité des différents processus envisagés. Une recherche bien ciblée, les caractéristiques de population correctement déterminées, un échantillonnage représentatif de celles-ci, un recueil de données adapté à la population mais aussi aux objectifs de la recherche, un traitement correct et objectif des données offrent les meilleures garanties de validité et de fiabilité à l'entreprise.

Ainsi, il n'existe pas de règles fixes pour déterminer la taille d'un échantillon. Il reste néanmoins certaines règles statistiques. L'usage à visée d'inférence de la statistique n'autorise pas un échantillon d'un effectif inférieur à 5. On parle souvent du chiffre « 30 », il s'agit en réalité, comme nous le verrons, d'une borne délimitant l'usage d'une loi statistique plutôt qu'une autre.

Les erreurs d'échantillonnage :

La qualité des résultats d'une enquête, la qualité de l'analyse comme des conclusions[22] sont fortement liées à la composition de l'échantillon. Lorsque l'échantillon est biaisé, la généralisation n'est pas possible car illégitime. Quels sont les biais générés par l'échantillonnage ?

- Les absences :

[22] On peut parler d'un effet « boule de neige ».

Une absence peut être temporaire ou définitive. Elle concerne un individu tiré au sort pour faire partie de l'échantillon mais absent lors de l'enquête. En cas d'absence définitive, la personne peut être remplacée en procédant à un nouveau tirage au sort dans la même catégorie de variables. L'échantillon peut également être corrigé en comparant les réponses obtenues par les personnes qu'on a pu joindre tout de suite à celles des absents temporaires.

- Les refus :

  Lorsque ce pourcentage est inférieur à 20 % et si les causes invoquées sont diverses et diversifiées, il est conseillé de remplacer en procédant de la même manière que pour un absent définitif. Au-delà de ce pourcentage, et surtout si la même cause de refus est évoquée, la question de la validité de l'enquête se pose. En effet, tout porte à penser que les réponses des non-répondants, par refus, seraient différentes de celles des répondants. Dans ce cas, il n'est pas nécessaire d'augmenter l'échantillon, le biais ne serait qu'accru. Pour éviter cet écueil, difficilement interprétable, un certain nombre de précautions sont à prendre dans la production de l'outil de recueil[23].

Echantillons indépendants et échantillons appareillés :

- Echantillons indépendants :

  Deux échantillons sont indépendants s'ils sont tous les deux représentatifs de la population parente, constitués au hasard et qu'une modification dans l'un des groupes n'a pas d'influence sur l'autre groupe. Il est impératif que le fait d'être élu dans un groupe soit le fruit du hasard.

  Exemple : patients opérés de PTH : G1 = EVA – G2 = ENA – (G1 ≠ G2).

- Echantillons appareillés :

  Deux échantillons sont appareillés lorsque chaque élément du premier est lié à un élément du second par une même relation. C'est également le cas d'une comparaison d'un pré-test à un post-test avec les mêmes sujets.

  Exemple : patients opérés de PTH : G1(T1) = EVA – G2 (T2) = ENA – (G1 = G2).

  Deux échantillons peuvent être appareillés même s'il ne s'agit pas des mêmes individus qui composent les deux échantillons. Il suffit que les sujets du premier aient le même score ou la même valeur que ceux du second.

  Exemple : deux stratégies d'éducation à la santé, vous appareillez les deux groupes sur une variable, celle du Q.I.. Pour chaque sujet de G1 ayant un QI = x, on fait correspondre un sujet de G2 ayant le même Q.I..

[23] Renvoi vers la balise *sigma*.

Chapitre 7 : Tableau de répartition.

Deux méthodes existent et coexistent, parfois, dans la présentation de résultats d'enquête : le tableau et le graphique. C'est sans doute par eux que la statistique devient statistiques. D'un point de vue pratique, l'élaboration judicieuse[24] du tableau de répartition permet la réalisation graphique de la distribution statistique étudiée.

Une manière habituelle de « faire des statistiques » consiste à calculer la fréquence d'apparition d'un caractère d'une variable ou d'une catégorie de réponse pour, ensuite, repérer sa répartition sur l'ensemble des caractères de la variable étudiée. A partir de ce point, les exemples seront bien souvent, comme les exercices, la meilleure façon de procéder.

1. Tableau de fréquence à un caractère qualitatif :

   Le codage proposé est bien souvent arbitraire, au mieux aléatoire. Le 1 peut correspondre avec la fréquence la plus élevé : le codage n'a-t-il pas eu lieu *a posteriori* ? Le codage reprend-t-il l'ordre des questions utilisé ? L'ordre des questions et/ou des codes a-t-il été influencé par la recherche conceptuelle ? Comment ?

Motif d'entrée SUS	Code	Effectif	Effectif cumulé	Fréquence
Malaise sur la voie publique	1	18		
Douleur thoracique	2	24		
Accident de la route	3	36		
Angoisses	4	12		
Maux de tête	5	09		
Douleurs abdominales	6	27		
TOTAL			TOTAL	1

2. Tableau de fréquence à un caractère quantitatif continu :

Niveau de dépendance	Effectif	Effectif cumulé croissant	Fréquence
0	0		
1	2		
2	6		
3	23		
4	11		
5	32		
6	18		
7	12		
8	5		
9	7		
10	4		
TOTAL		TOTAL	

[24] y compris à l'aide d'un tableur informatique, comme Excel©.

3. Tableau de fréquence à deux caractères qualitatifs :

   Lorsqu'une étude s'intéresse à deux caractères différents ou deux variables, il est possible de présenter l'ensemble des observations dans un tableau à double entrée ou de contingence.

   Exemple : Deux variables (sexe / petit-déjeuner), chaque variable présente deux modalités.

	Filles	Garçons	Total
Alimentation	20 (0,69₀)	35 (0,7₀₀)	55 (0,696)
Absence aliment.	9	15	24
TOTAL	29	50	79

   Conclusion : la variable sexe n'est pas discriminatoire.

4. Pourcentage :

   Le pourcentage reste avec la moyenne que nous étudierons, l'expression statistique la plus commune. Pourtant, son utilisation se doit d'être nuancée. Le pourcentage est une fréquence calculée sur un effectif total supérieur ou proche de 100 et, de ce fait, ramené à 100.

   Exemple : Deux variables (sexe / petit-déjeuner), chaque variable présente deux modalités.

	Filles	Garçons	Total
Alimentation	118 (63,44 %)	335 (80,92 %)	453 (75,50 %)
Absence aliment.	68	79	147
TOTAL	186	414	600

5. Décimales :

   En la matière, il existe aussi des conventions qui justifient y compris par l'arithmétique. Le nombre de chiffres après la virgule se doit d'être homogénéisée au moins dans le même tableau ou graphique. En outre, plus fondamentalement d'ailleurs, la règle pour conserver ces décimales doit être standardisée en vous rappelant les deux possibilités : l'arrondi ou la troncature.

Chapitre 8 : Représentations graphiques.

La représentation graphique permet d'avoir une vue d'ensemble de la répartition des caractères observés, de repérer les caractéristiques essentielles et de comparer des séries différentes. Vous pouvez donc représenter une distribution par des graphiques. Nous répéterons la règle suivante : l'objectif est d'illustrer le propos et non de s'y substituer[25].

Avec quelques chiffres élaborés après un périple statistique : une moyenne, un écart type, une probabilité, un intervalle de confiance, le graphique reste la partie la plus visible de l'exercice. Il sera aussi l'outil électif, et non exclusif, lors de leur communication. L'emprise visuelle et l'interprétation qui découle de ce sens, témoignent, une fois de plus, de la prudence capitale dans son emploi. Désormais informés, qu'en statistiques comme en comptabilité, la manipulation[26] prend le pas de la négligence ou de l'erreur.

Identiquement à d'autres manœuvres statistiques, il est crucial de repérer les caractéristiques des variables utilisées puisqu'elles conditionnent, à leur tour, le type de représentations graphiques des résultats. Comme d'accoutumé, un certain nombre de conventions sont présentes et sont à respecter : un titre, des légendes, l'échelle de mesure, etc.. Ici aussi, des choix sont réclamés et à pondérer en fonction de l'objectif poursuivi[27] et pareillement, la modération est souvent préférable à l'excès. Les moyens de visualisation, proposés par Excel®, sont trop diversifiés pour être utilisés de manière intégrale. De manière transversale, enfin, il faut veiller à l'homogénéité de présentation de vos graphiques et tableaux dans un souci de cohérence.

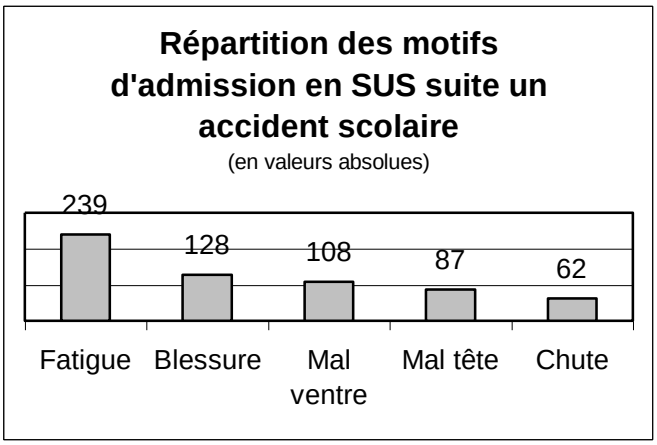
1. Caractères qualitatifs : diagrammes en barres ou circulaires :

   Une représentation en barre est constituée d'un ensemble de lignes ou de rectangles, dont la hauteur est proportionnelle aux effectifs, exprimés en valeurs absolues ou en proportion... Le diagramme circulaire est un cercle partagé en secteurs de surface proportionnelle aux effectifs.

Mal de tête	Mal au ventre	Blessure sportive	Fatigue	Chute
87	108	128	239	62
13,94 %	17,30 %	20,51 %	38,30 %	9,94 %

   Graphique en page suivante :

[25] Le cours de « communication didactique » s'enquiert de ces considérations.
[26] Nous ferons quelques mises en situation dans le chapitre 11 de cette balise.
[27] Notion récurrente de contribution.

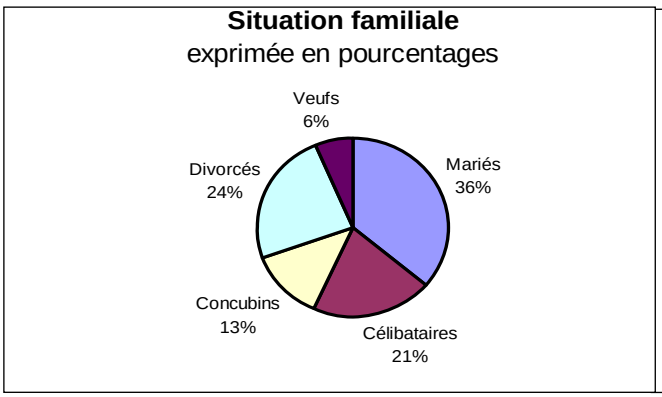


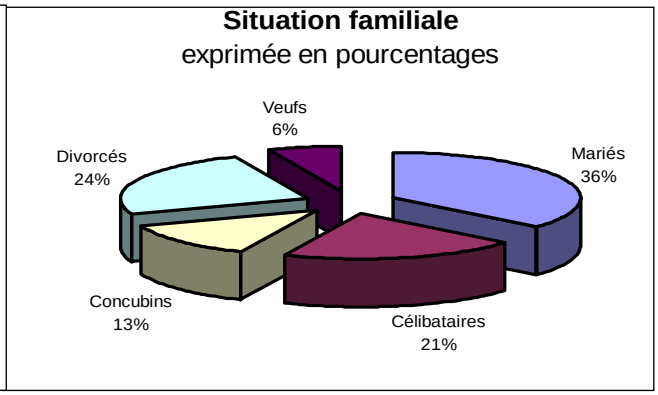
Une représentation graphique circulaire à secteurs, appelée communément « camembert », donne ce genre de visualisation. La prudence[28] sera une nouvelle fois soulignée dans l'usage de certains aspects visuels visant à, consciemment ou inconsciemment, tronquer le discours « statistique ». Sachez déjà que la représentation en 3D comme le second donne des perspectives qui fausse la vision, surtout si on tourne et éclate le disque afin de placer de front l'item qui doit focaliser l'attention des spectateurs-auditeurs.

Exemple : situation familiale des enquêtés

Mariés	Célibataires	Concubins	Divorcés	Veufs
356	204	128	239	62

Le camembert classique puis le même en 3D éclaté :





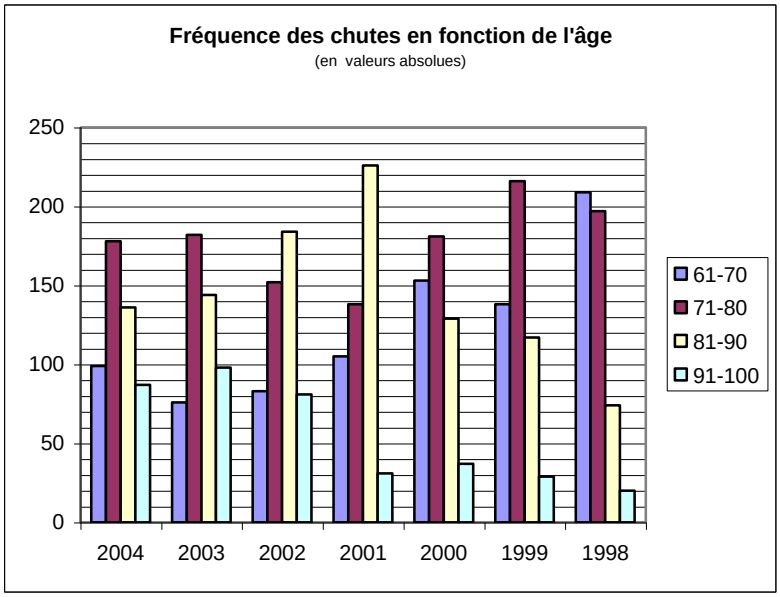
Passerelle Q/q : les données peuvent parfois être traitées de façon plus détaillée. L'usage le plus courant est de croiser deux échelles comme dans l'exemple suivant. L'objectif de la démarche de recherche est de confronter certaines données entre elles afin de vérifier un certain nombre d'hypothèses préexistant à l'enquête.

[28] Est-ce un synonyme de *méthodologie* ?

Exemple : chutes de patients hospitalisés répertoriés en fonction de l'âge.

Années	61 à 70 ans	71 à 80 ans	81 à 90 ans	91 à 100 ans
2004	99	178	136	87
2003	76	182	144	98
2002 - …	…	…	…	…

Représentation graphique de la courbe de distribution.



Il faut garder aussi en mémoire que l'excès d'information tue l'information ; un graphique ou un tableau trop chargé rebute souvent le lecteur. Tous ces moyens, au même titre que les outils, doivent être pondérés en fonction du ou des objectifs poursuivis[29].

2. Caractères quantitatifs discontinus : diagrammes en bâtons :

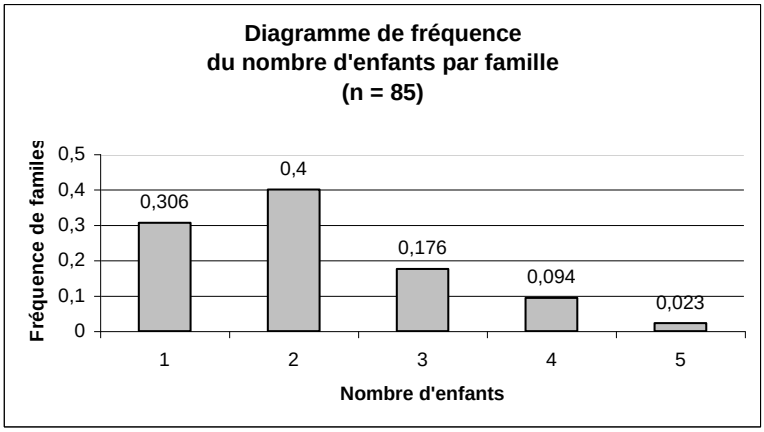
Il s'agit de présenter la fréquence plutôt que l'effectif, la fréquence est placée en ordonnée. La longueur du bâton est proportionnelle à la fréquence observée. Ici, il est nécessaire de donner la taille de l'effectif global : celui de l'échantillon considéré.

Exemple : nombre d'enfants par famille enquêtée :

Nombre d'enfants	1	2	3	4	5	Total
Effectif	26	34	15	8	2	85
Fréquence	0,306	0,400	0,176	0,094	0,023	

Graphique en page suivante :

[29] Voir, à ce propos, l'exemple des « motifs d'admission aux urgences pour accident scolaire » qui ne présente certainement pas tous les motifs d'admission. D'autre part, le classement peut être modifier lors du traitement.



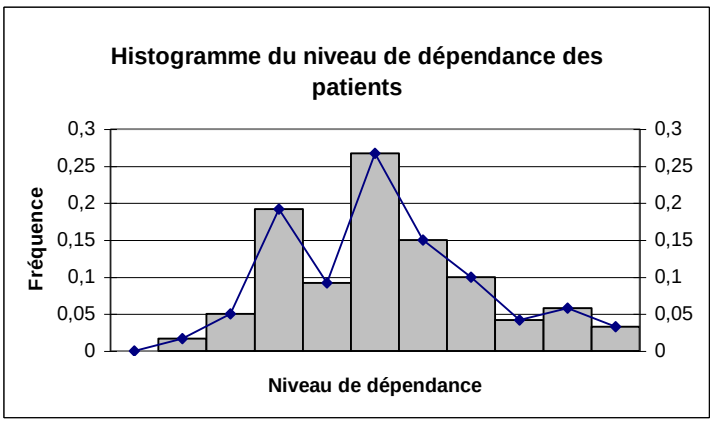
3. Caractères quantitatifs continus : l'histogramme :

   Pour construire ce type de graphique, chaque classe est représentée par un rectangle dont la base est égale à l'intervalle de classe et la hauteur à l'effectif correspondant. La base comme la hauteur peuvent alors varier, on parle d'« aire ».

   Exemple : niveau de dépendance de personnes hospitalisées.

Niveau de dépendance	E[30]	ECC	Fr	FrCC
0	0			
1	2			
2	6			
3	23			
4	11			
5	32			
6	18			
7	12			
8	5			
9	7			
10	4			**1**

   Représentation graphique :



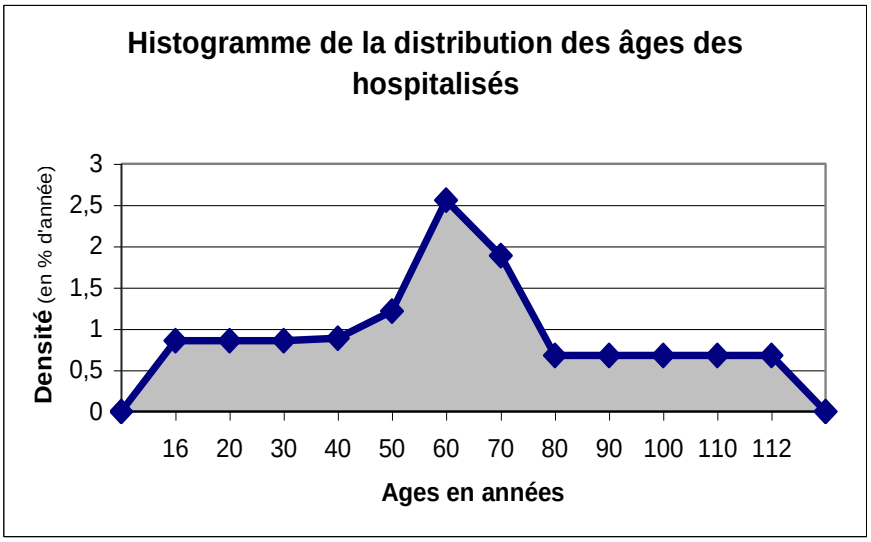
   La ligne qui relie les points définit le polygone des fréquences.

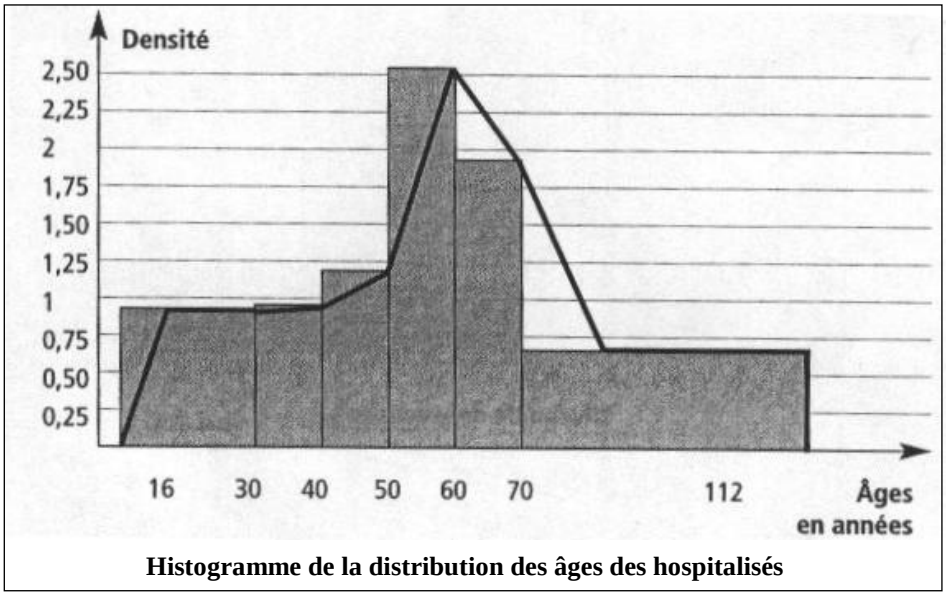
[30] Légende : E pour effectif, ECC pour effectif cumulé croissant, Fr pour fréquence, FrCC = fréquence cumulée croissante

Exemple bis : Distribution des âges des personnes hospitalisées

Classes d'âges	Effectif	Effectif cumulé croissant	Effectif en %	Amplitude d'années	Densité en % d'années
16 à 30 ans	39				
31 à 40 ans	27				
41 à 50 ans	35				
51 à 60 ans	75				
61 à 70 ans	53				
71 à 112 ans	91				

Lorsque l'amplitude des classes est variable, il essentiel de calculer la densité des classes. L'amplitude correspond au nombre d'années concernant chaque classe. La densité en pourcentage se calcule en divisant l'effectif en pourcentage d'une classe par l'amplitude de cette classe. L'histogramme est alors représenté de même que le polygone reliant les sous-classes[31].





Histogramme de la distribution des âges des hospitalisés

[31] Dans l'exemple par tranche de 10 ans, sauf aux extrémités.

Chapitre 9 : Les indices de résumé statistique.

Premières opérations mathématiques, ces indices permettent de résumer les données décrivant une population, un échantillon. Ils représentent aussi. Les indices de tendance centrale informent sur les valeurs dominantes tandis que les indices de dispersion renseignent sur l'étalement des données autour des valeurs centrales. Ils déboucheront par la suite sur des lois, en particulier celle de Gauss, sur lesquelles des tests statistiques se basent.

1. Les indices de tendance centrale :

1°- Le mode :

Le mode correspond à la catégorie de réponse où la valeur d'un caractère est la plus représentée, ou à la fréquence maximale. Il se calcule sur toutes les échelles. La classe comportant l'effectif le plus important est nommée classe modale. Une distribution peut aussi être bi- voire pluri-modale.

Exemple : dans le graphique intitulé « diagramme du nombre d'enfants par famille », déterminez la classe modale. Le mode ou classe modale de cette distribution est ..... .

2°- La médiane ou rang médian :

La médiane est la valeur de variable qui divise l'effectif total en deux parties égales. Pour un ensemble discontinu de *n* valeurs, elle est la valeur centrale de l'ensemble si *n* est impair ou la moyenne des deux valeurs centrales si *n* est pair.

Sur une échelle ordinale ou numérique, elle nécessite le calcul de l'effectif cumulé, croissant ou décroissant. Si le nombre total d'observations *N* est pair, la médiane est de N/2. Si le nombre total d'observations *N* est impair, la médiane est de N/2+ ½.

Exemple n° 1 : niveau de douleur à l'admission suite à un …

Niveau de douleur	Effectif	ECC
0	9	
1	17	
2	25	
3	31	
4	12	
5	45	

- → La classe modale de cette distribution est ..... ,
- → La médiane est calculé comme suit :
- → Le rang médian de cette distribution est à ..... .

Sur des variables continues, et donc un histogramme, le mode est calculé sur la densité et la médiane doit être précisée.

Exemple n° 2 : Distribution des âges des personnes hospitalisées.

Classes d'âges	Effectif	Effectif cumulé croissant	Effectif en %	Amplitude d'années	Densité en % d'années
16 à 30 ans	39				
31 à 40 ans	27				
41 à 50 ans	35				
51 à 60 ans	75				
61 à 70 ans	53				
71 à 112 ans	91				

→ La classe modale est alors la classe des ..... ans.

→ La médiane se calcule comme suit :

→ Le rang médian se trouve dans la classe des ..... ans.

Le calcul doit être affiné. Une classe détermine un intervalle au sein duquel une dispersion est aussi présente.

La médiane ($M_é$) se calcule selon la formule suivante :

$$M_é = (X_{i-1}) + A_i * \frac{N/2 - (NCCi - 1)}{N_i}$$

Procédure :

Rassembler les « ingrédients » avant de calculer ;

- Effectif total de la population : N = ...
- Effectif de la classe concernée : N_i = ...
- Effectif cumulé croissant de la classe concernée : NCC_i = ...
- Effectif cumulé de la classe concernée – 1 : (NCC_{i-1}) = ...
- Valeur de la borne inférieure de la classe : (X_{i-1}) = ...
- Amplitude de la classe concernée : A_i = ...

Pour l'exemple :

Calcul de $M_é$ = ...

La médiane de l'histogramme est à ..... ans.

3°- La moyenne arithmétique :

Elle ne se calcule que sur une échelle numérique et détermine le centre de gravité d'une distribution. Elle peut s'écrire M_x , m ou $\overline{X}$. Sa formule est :

$$M_x = \frac{\sum n_i x_i}{N}$$

- x_i désigne la valeur de la classe $_i$ si vous n'avez pas effectué de regroupement et la valeur centrale de cette même classe si vous avez effectué un regroupement.
- n_i définit l'effectif de la classe.
- N signifie l'effectif de la population ou de l'échantillon (n).

Il s'agit donc de calculer, dans un premier temps, le produit n_i x_i pour chaque classe, d'effectuer la somme de ces produits avant de la diviser par l'effectif total. Reprenons l'exemple utilisant les niveaux de dépendance.

Exemple : niveau de dépendance de personnes hospitalisées

Niveau de dépendance	E			
0	0			
1	2			
2	6			
3	23			
4	11			
5	32			
6	18			
7	12			
8	5			
9	7			
10	4			

→ Le mode est ..... ;

→ La médiane est ..... ;

→ La moyenne arithmétique est ..... .

2. Les indices de dispersion :

Ces indices nous renseignent sur l'étalement des données, leur répartition autour de la moyenne. Elles peuvent être plus ou moins étalées. L'objectif de ces indices est de décrire cette dispersion.

1°- L'intervalle de variation[32] :

Il indique l'étendue du positionnement des données en relevant la différence entre la plus grande et la plus petite des valeurs exprimées de la variable : les extrêmes.

2°- L'écart est la valeur absolue de la différence de deux valeurs de la variable.

3°- L'écart moyen :

L'écart moyen correspond à la moyenne arithmétique des valeurs absolues de l'écart de chaque valeur à la moyenne arithmétique de l'ensemble des données.

4°- L'écart moyen relatif :

Il indique le rapport entre l'écart moyen et la moyenne arithmétique des données.

5°- Les quartiles :

Les quartiles correspondent aux trois valeurs qui séparent la distribution en quatre parties égales et proportionnelles au nombre d'observations. Le premier quartile est le quartile inférieur, le second correspond à la médiane et le troisième est dit quartile supérieur. Ils se calculent sur une échelle ordinale ou numérique. L'écart interquartile est la distance entre le

[32] Cet intervalle est désigné par le symbole W.

premier et le troisième quartile. On parle également de percentile ; il s'agit alors de diviser la distribution en 100 parties, soit de 1 %. Il existe différents types de *ntiles* selon le niveau de division choisi par le chercheur.

Exemple n° 1 : enquête sur les lombalgies dont mesure de la douleur par une échelle discontinue. La variable « sexe » est introduite. La population totale s'élève à 435 individus.

Hommes	0	1	2	3	4	5	6	7	8	9	10
Effectif	6	12	25	12	48	34	27	52	46	29	7
ECC											

- → Le mode est ..... ;
- → La médiane est ..... ;
- → La moyenne est égale à ..... ;
- → L'intervalle de variation : ..... ;
- → Le quartile inférieur : ..... et sa valeur associée est ..... ;
- → Le quartile supérieur est à ..... et sa valeur associée est .....;
- → L'écart interquartile est dès lors de ..... ;
- → L'écart moyen se calcule à l'aide d'un tableau de conversion.

Valeur x_i	$\lvert x_i - m \rvert$	$n_i \ (x_i - m)$
0		
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
Total		

- → L'écart moyen est égal à ..... ;
- → L'écart moyen relatif est de : ..... .

Femmes	0	1	2	3	4	5	6	7	8	9	10
Effectif	0	9	2	12	31	25	36	12	10	0	0

- → Le mode est ..... ;
- → Le rang médian est ..... ;
- → La moyenne est égale à ..... ;
- → L'intervalle de variation : ..... ;
- → Le quartile inférieur : ..... et sa valeur associée est ..... ; ;
- → Le quartile supérieur est à ..... et sa valeur associée est ..... ;
- → L'écart interquartile est de ..... ;

→ L'écart moyen est égal à ..... ;

→ L'écart moyen relatif est de ..... .

Conclusion de l'exemple n° 1 : en comparant les deux écarts moyens relatifs, on constate que les données sont moins dispersées dans la population/l'échantillon ..... que dans celle/celui des ..... .

Lorsque la démarche s'applique sur des intervalles, le calcul s'opère alors sur la densité en considérant le centre de la classe.

Exemple n° 2 : distribution des âges des hospitalisés.

Classes d'âges	Effectif	Effectif en %	Centre de la classe x_i	$\lvert x_i - m_i \rvert$	Densité en $n_i * \lvert x_i - m_i \rvert$
16 à 30 ans	39				
31 à 40 ans	27				
41 à 50 ans	35				
51 à 60 ans	75				
61 à 70 ans	53				
71 à 112 ans	91				
TOTAL					

→ La moyenne ..... ans.

→ L'écart moyen est donc de ..... ans.

Conclusion de l'exemple n° 2 : Les observations s'écartent de ..... ans, l'âge moyen, de ..... ans.

6°- La <u>variance</u> :

La variance d'une série statistique, désignée par le symbole V, est la moyenne de la somme des carrés des écarts par rapport à la moyenne. Elle permet de calculer un indice de dispersion à partir des écarts des différentes valeurs observées par rapport à leur moyenne. Dans une distribution étalée, les écarts sont plus importants. Sa formule est la suivante :

$$Vx = \frac{\Sigma\, n_i\, (x_i - m)^2}{N}$$

Exemple des lombalgies :

1°- la population masculine :

x_i	n			
0	6			
1	12			
2	25			
3	12			
4	48			
5	34			
6	27			
7	52			
8	46			
9	29			
10	7			

→ La variance « masculine », soit Vx, est de ..... .

2°- la population féminine :

y_i	n			
0	0			
1	9			
2	2			
3	12			
4	31			
5	25			
6	36			
7	12			
8	10			
9	0			
10	0			

→ La variance « féminine », soit Vy, est de ..... .

Conclusion de l'exemple : au sein des échantillons convoqués, la variance de distribution observée chez les ..... est nettement ..... à celle observée chez les ..... .

7°- L'écart type :

Il s'obtient en calculant la moyenne quadratique des écarts des valeurs d'un caractère à leur moyenne arithmétique, c'est, autrement dit[33], la racine carrée de la variance. L'écart type est symbolisé par σ ou s. Il est le plus significatif des indices de dispersion. L'écart type s'exprime dans la même unité que celle de la modalité étudiée ou des données.

La dispersion est d'autant plus faible que l'écart type est plus petit.

$$\text{Formule : } \sigma = \sqrt{Vx}$$

Exemple des lombalgies :

→ population masculine : Ecart type : σ_x = ... ;

→ population féminine : Ecart type : σ_y = ...

Prémices d'interprétation :

La statistique n'interprète pas, nous le savons, elle est outil d'analyse et de production de statistiques. Nous possédons désormais les deux indices permettant de décrire, par deux valeurs, la distribution d'une ou plusieurs variables. Fréquemment, des résultats sont présentés, dotés isolément de leur moyenne. En espérant qu'elle soit arithmétique, à elle seule, elle ne peut être la représentation d'une enquête. L'un n'allant pas sans l'autre, l'écart-type donne le niveau de validité de votre moyenne et donc de votre enquête. Sans son indice de dispersion, il y a lieu de douter sur la pertinence de la moyenne annoncée. Nous le verrons avec Gauss, l'écart-type relativise la moyenne, qui, elle, ne peut rester seule !

[33] et plus facile à retenir, aussi !

8°- L'écart type et la courbe de Gauss :

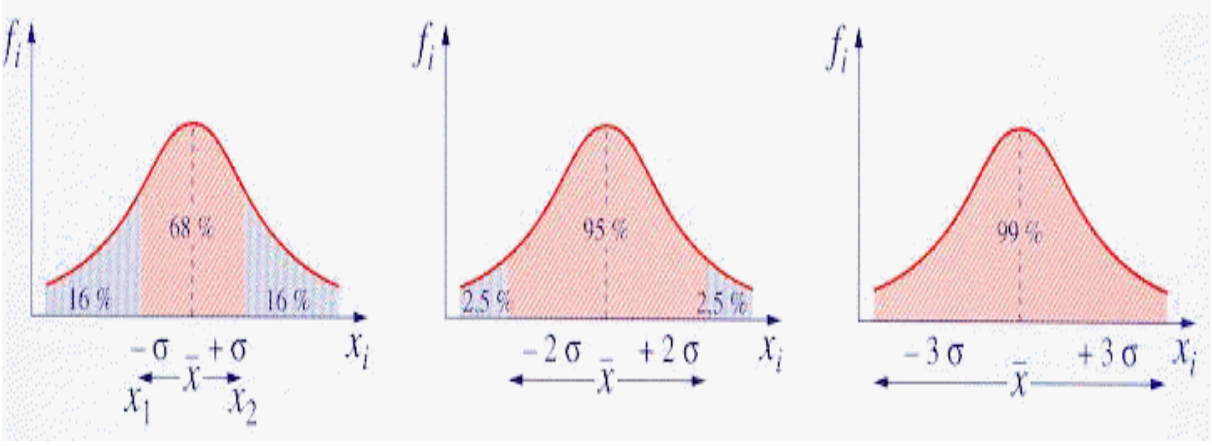
Connue, au moins dans sa forme, la courbe de Gauss reste un outil fondamental, bien que, par certains, contestée. La courbe de Gauss ou « *curve bell* », n'est en réalité que la représentation graphique d'une distribution jugée normale car fréquemment remarquée. Une loi et une table permettent de s'y référer en vue de tester les hypothèses ou d'estimer la moyenne vraie. Ces opérations statistiques seront traitées dans un chapitre à venir.

La courbe de Gauss est l'expression graphique d'une distribution qui correspond à la loi « normale ». Cette loi comme d'autres sont d'utilité statistique.

Quelles sont les propriétés de cette répartition ? :

- les valeurs du mode, de la médiane et de la moyenne sont égales,
- la courbe est symétrique,
- 68 %[34] des valeurs sont comprises entre la moyenne moins un écart type et la moyenne plus un écart type,
- 95 % des valeurs sont comprises entre la moyenne moins deux écarts types et la moyenne plus deux écarts types,
- 99,75 % des valeurs sont comprises entre la moyenne moins trois écarts types et la moyenne plus trois écarts types

Cette distribution est dite théorique ou « *normale* » car elle est la distribution de probabilité la plus courante. Ces intervalles sont nommés intervalles de confiance. Ces deux termes, probabilité et intervalle de confiance, reviendront ultérieurement dans cet exposé ; ils renvoient aux opérations statistiques mentionnées ci-dessus.



[34] Ces chiffres peuvent varier d'une décimale.

Chapitre 10 : Etude de corrélations.

Jusqu'à présent, nous nous sommes intéressés à des questions étudiant le comportement statistique d'une seule variable. Il existe cependant toute une gamme de problèmes statistiques où l'on s'intéresse à la relation, la co-relation entre deux variables statistiques[35], une dépendance fonctionnelle au sein d'une série double, une interdépendance. En statistique, les corrélations naissent de l'analyse de régression.

Exemples de questions :

- les individus les plus grands sont-ils les plus lourds ?
- le revenu d'une famille a-t-il une influence sur les résultats scolaires des enfants ?
- y a-t-il une relation entre le tabagisme et les cancers du poumon ?
- le rendement en céréales dépend-il de la quantité d'engrais utilisée ?
- la productivité d'une entreprise est-elle liée au salaire des ouvriers ou employés ?

Dans ces questions, nous désirons savoir si le comportement d'une variable est influencé par la valeur d'une autre variable :

Taille	⟺	Poids
Revenu familial	⟺	Résultats scolaires
Tabagisme	⟺	Cancers pulmonaires
Rendement	⟺	Engrais
Productivité	⟺	Salaires

Si c'est le cas, il devient possible de :

- contrôler le second si on sait régler le premier,
- prévoir le second si on n'observe plus que le premier.

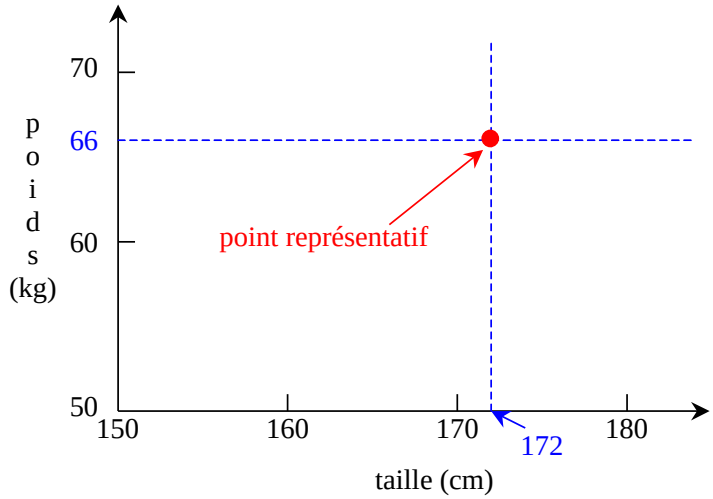
Ces deux opérations seront d'autant plus précises que l'interdépendance est forte.

La relation peut être causale ou non.

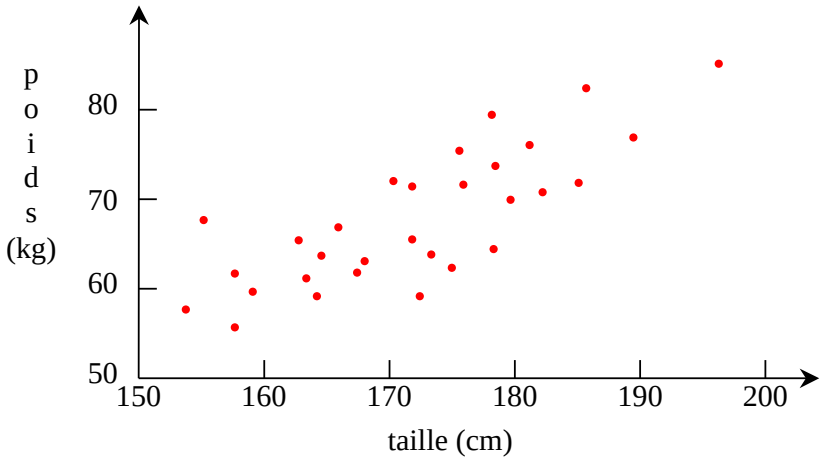
1. Nuage de points :

Pour étudier les relations ou corrélations entre deux variables, il est possible de porter chaque mesure sur un graphique. Le graphique ainsi constitué est appelé diagramme de régression et prend la forme du nuage de points.

[35] Tout particulièrement dans la méthode différentielle.



Dans l'exemple, est porté sur le graphique, pour chaque individu de l'échantillon : sa taille en abscisse[36] et son poids en ordonnée.



Relation entre le poids et la taille dans un échantillon de 30 individus.

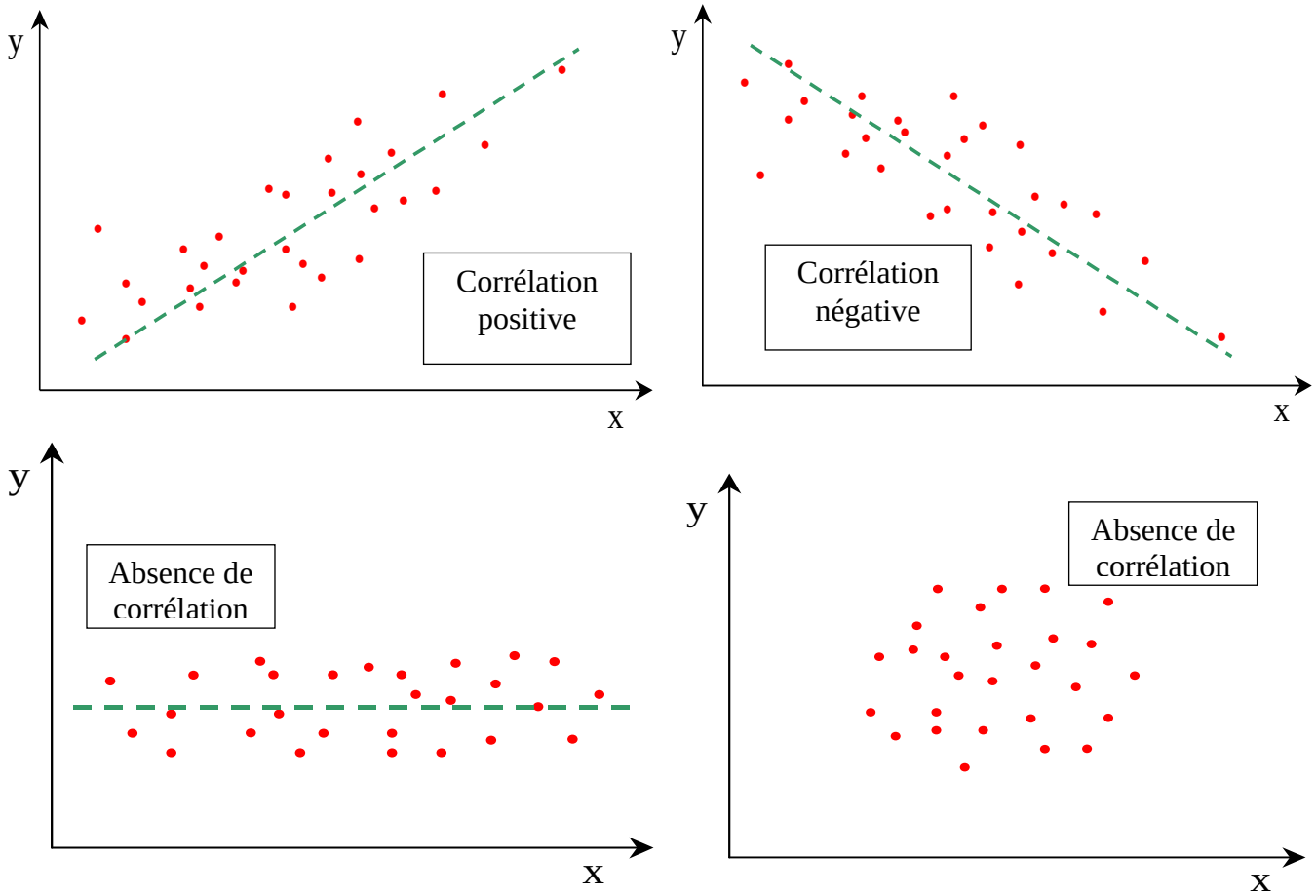
Chaque individu est donc, dans ce graphique, représenté par un point, un point représentatif. Il y aura donc autant de points qu'il y a d'individus dans l'échantillon.

2. Droite de régression :

On peut, par la pensée ou réellement, tracer une droite qui passe au mieux par ces points ; c'est-à-dire au milieu du nuage de points. Cette droite de régression témoigne de la linéarité de relation entre les variables étudiées. Nous sommes faces à des régressions linéaires et monotones. Certaines courbes de régression peuvent être non linéaires comme en témoigne la figure *f* du tableau en page 33. Dès lors, le nuage peut prendre différentes formes : parabole, hyperbole, sinusoïde…. Il existe d'autres régressions que linéaires ; nous les étudierons pas.

[36] L'abscisse d'un point correspond à sa projection sur l'axe horizontal Qu'est-ce alors que l'ordonnée ?

- Si cette droite « monte », on dira qu'il y a corrélation positive entre les deux variables ;
- Si elle « descend », la corrélation est négative ;
- Si elle est « horizontale » ou si on ne peut pas décider, il y a absence de corrélation[37].



Une relation monotone est déclarée positive lorsque les caractères étudiés varient dans le même sens, c'est-à-dire que :

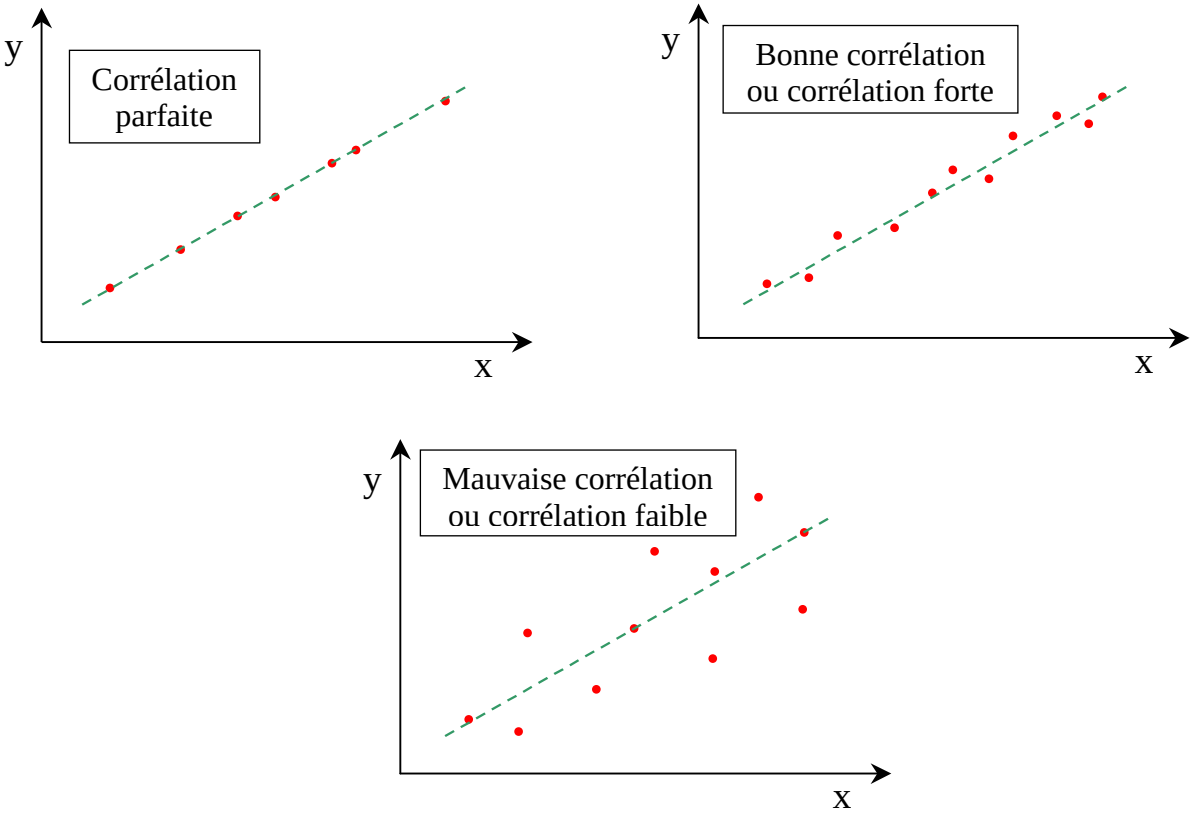
- les valeurs fortes de X correspondent généralement aux valeurs fortes de Y ;
- les valeurs moyennes de X correspondent généralement aux valeurs moyennes de Y ;
- les valeurs faibles du X correspondent généralement aux valeurs faibles de Y.

Une relation monotone est déclarée négative lorsque les caractères étudiés varient en sens inverse, c'est-à-dire que :

- les valeurs fortes de X correspondent généralement aux valeurs faibles de Y ;
- les valeurs moyennes de X correspondent généralement aux valeurs moyennes de Y ;
- les valeurs faibles du caractère X correspondent généralement aux valeurs fortes de Y.

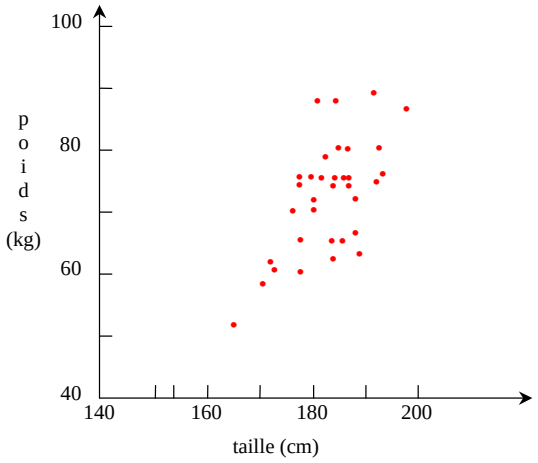
[37] Pour être complet, il faut ajouter « linéaire et monotone ».

En seconde intention, après avoir déterminer le sens de cette relation, il est possible d'approcher la qualité de la corrélation entre deux variables ; elle se mesure par la dispersion des points autour de la relation moyenne, c'est-à-dire autour de la droite de régression.



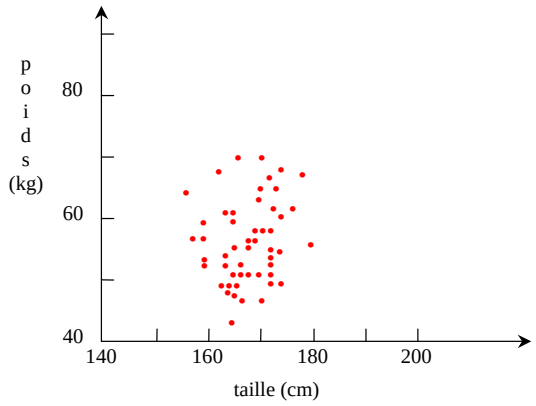
Exemples :

1. Corrélation entre le poids et la taille pour les garçons de 3ème cadres (2006).



On constate une augmentation du poids avec la taille, il s'agit d'une corrélation positive : les garçons les plus grands sont généralement les plus lourds. Mais la dispersion des points est assez grande, donc la corrélation est assez faible.

2. Corrélation entre le poids et la taille pour les filles de 3ème cadres (2006).



On ne constate pas de relation entre le poids et la taille, le poids des filles est indépendant de leur taille : il y a absence de corrélation.

3. Méthode des moindres carrés :

Si on se contente de tracer à main levée la droite qui passe au mieux par les points représentatifs, différentes personnes vont obtenir des résultats différents. Pour pallier cette subjectivité, il existe une équation afin de déterminer la meilleure droite : c'est la méthode des moindres carrés. Ces calculs, comme le tracé aléatoire de la droite, peuvent être remplacés par le calcul du coefficient de corrélation.

4. Coefficient de corrélation :

Sens et qualité d'une corrélation peuvent être, aisément, mesurés par le coefficient de corrélation *r* dont suit la formule : $r = \frac{\sum (X - \overline{X}).(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2} \times \sqrt{\sum (Y - \overline{Y})^2}}$

Ce coefficient peut aussi s'écrire dans cette formule où l'on retrouve les écarts-types pour chaque variable en dénominateur et la covariance en numérateur. La covariance s'appuie sur le calcul de la moyenne pour chaque variable.

$$r = \frac{\sigma_{xy}}{\sigma_x * \sigma_y}$$

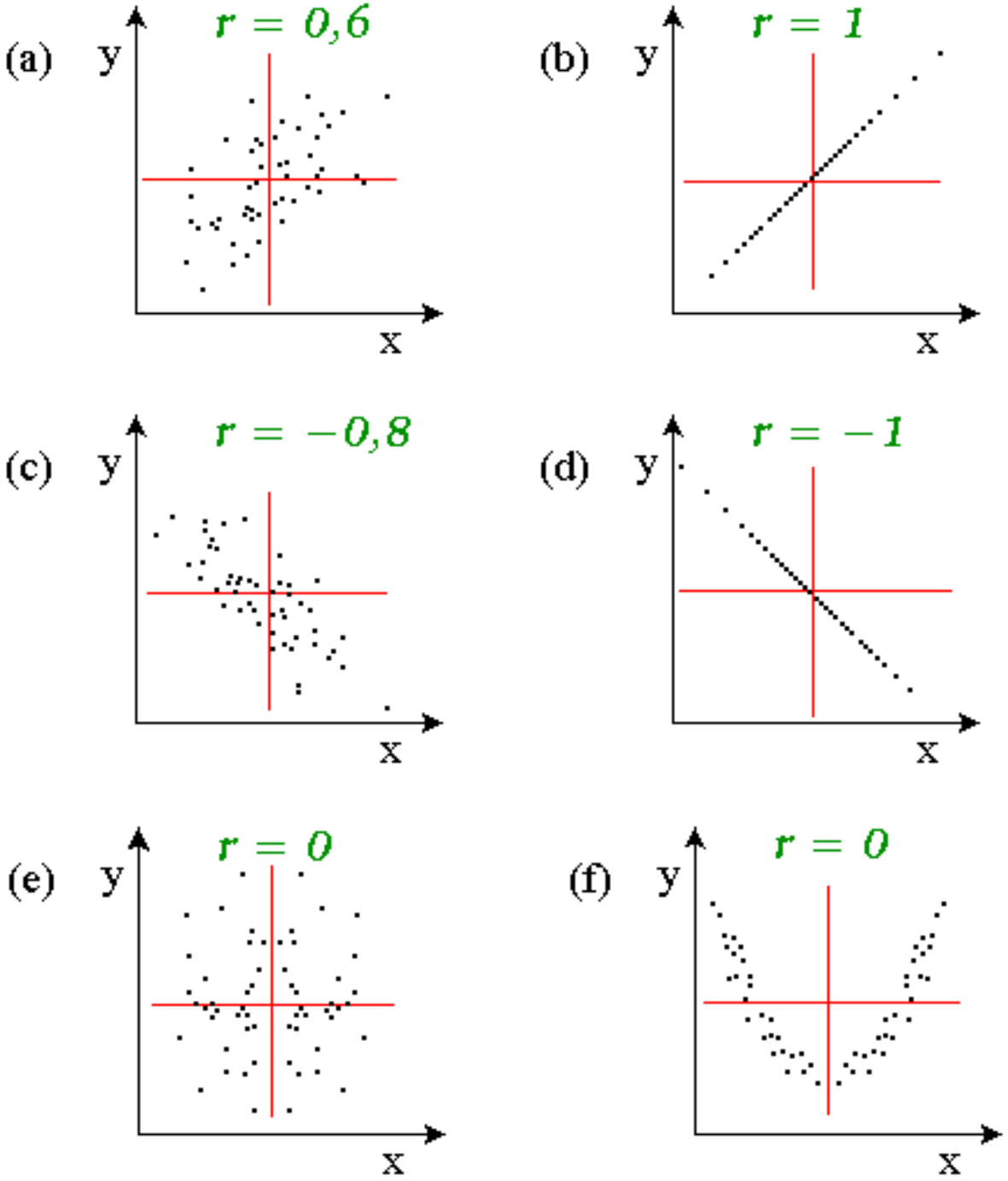
Ce coefficient *r* est appelé aussi coefficient de Bravais Pearson. Il permet donc d'indiquer le degré de liaison linéaire entre deux variables. Le coefficient de corrélation est compris entre –1 et +1. Ce nombre sans unité traduit donc la plus ou moins grande dépendance linéaire entre deux caractères. Plus il s'éloigne de zéro, meilleure est la corrélation : la corrélation est forte.

$r = +1$	corrélation positive parfaite
$r = -1$	corrélation négative parfaite
$r = 0$	absence totale de corrélation

Le signe de la valeur indique donc le sens de la relation tandis que la valeur absolue indique l'intensité de la relation. Autrement dit, il indique la capacité à prédire les valeurs d'un caractère en fonction de celles de l'autre.

positif	Les deux variables évoluent dans le même sens.
négatif	La relation entre les deux variables est inversée.

Quelques exemples de corrélations vont sont proposés avant de passer à un exercice.



Exemple n° 1 :

Supposons un échantillon aléatoire de quatre firmes pharmaceutiques présentant leurs dépenses de recherche (X) et leurs profits (Y), exprimés en millions de dollars :

Dépenses (X)	Profits (Y)
40	50
40	60
30	40
50	50

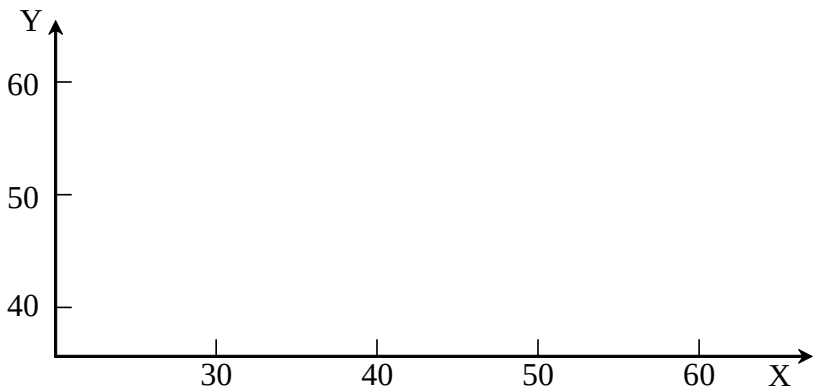
Calculons tout d'abord X et $\overline{Y}$ puis utilisons un tableau de conversion :

X	Y					
40	50					
40	60					
30	40					
50	50					

Afin d'obtenir le coefficient de corrélation :

$$r = \frac{\sum\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sqrt{\sum\left(X - \overline{X}\right)^2} \times \sqrt{\sum\left(Y - \overline{Y}\right)^2}} =$$

La corrélation est ..... et de qualité ..... .



Exemple n° 1 :

La corrélation entre la taille et le poids pour les garçons de 3ème cadres donne :

r = 0,61 : la corrélation est donc positive, de qualité moyenne.

De la même manière, pour les filles, on obtient :

r = 0,20 : la corrélation est positive mais de très faible qualité.

5. Interprétation d'une corrélation :

Le coefficient de corrélation nous donne des informations sur l'existence d'une relation linéaire entre les deux grandeurs considérées. Un coefficient de corrélation nul ne signifie pas l'absence de toute relation entre les deux grandeurs. Il peut exister une relation non linéaire entre elles. Dans tous les cas, la connaissance de X nous donne des informations sur la valeur de Y.

Il ne faut pas confondre corrélation et relation causale ! Une « bonne », même très « bonne »[38] corrélation entre deux grandeurs peut révéler une relation de cause à effet entre elles, mais pas nécessairement.

Exemples :

1°- Si on compare la durée de vie des individus à la quantité de médicaments pour le cœur qu'ils ont absorbée, on observera probablement une « bonne » corrélation négative. Il serait imprudent de conclure que la prise de médicaments pour le cœur abrège la vie des individus…

En fait, dans ce cas, la corrélation est l'indice d'une cause commune : la maladie de cœur. Cette notion de cause commune est également appelée « facteur de confusion ».

2°- Le soleil tire son énergie de réactions nucléaires transformant l'hydrogène en hélium. Notre société tire une part substantielle de son énergie de la combustion du pétrole. Si on compare, année après année, la quantité d'hélium contenue dans le soleil au prix moyen du pétrole, on obtiendra une « bonne » corrélation positive, sans qu'il y ait la moindre relation de cause à effet, ni aucune cause commune.

3°- Depuis une dizaine d'années, la taille de mon fils, né en 1989, est très bien corrélée avec la puissance de calcul des ordinateurs personnels. Cette excellente corrélation ne révèle bien évidemment aucune relation de cause à effet, ni cause commune.

💣💣💣💣 L'existence d'une corrélation, aussi « bonne » soit elle, n'est jamais la preuve d'une relation de cause à effet. 💣💣💣💣

6. Table des valeurs de *r* et coefficient en rang de Spearman :

Il existe des tests statistiques qui s'appliquent aux corrélations. Nous les examinerons plus tard lorsque nous pourrons nous intéresser à ces techniques statistiques.

[38] plus justement : *parfaite*.

Chapitre 11 : Représentation graphique, communication et autres manœuvres…

Dans ce chapitre, nous passerons en revue un certain nombre de manœuvres qui peuvent être faites intentionnellement ou par négligence. Le but est d'aiguiser votre sagacité déjà dans vos lectures et ensuite, de vous aider à pondérer l'usage de cette communication visuelle. De la sorte, vous ne pourrez plus pêcher par négligence ou méconnaissance. Les chiffres ont de grands pouvoirs, leur manipulation en a tenté plus d'un … avec plus ou moins de succès !

Les exemples qui suivent ont pour but de comprendre l'erreur commise dans la représentation, et par-là l'interprétation erronée, biaisée que le lecteur en retirera. Le graphique sera, parfois, « relooké » et réinterprété correctement.

1. Regroupement par classes et graphiques :

Le regroupement en classes des différentes mesures nécessite l'application d'un intervalle, dont la représentation graphique sera attribuée par sa largeur. La largeur choisie les classes dépendra :

- de la finesse de la représentation désirée,
- de la taille de l'échantillon étudié.

Pour que la représentation ait suffisamment de précision, il faut que chaque classe contienne, en général, un nombre suffisant d'individus.

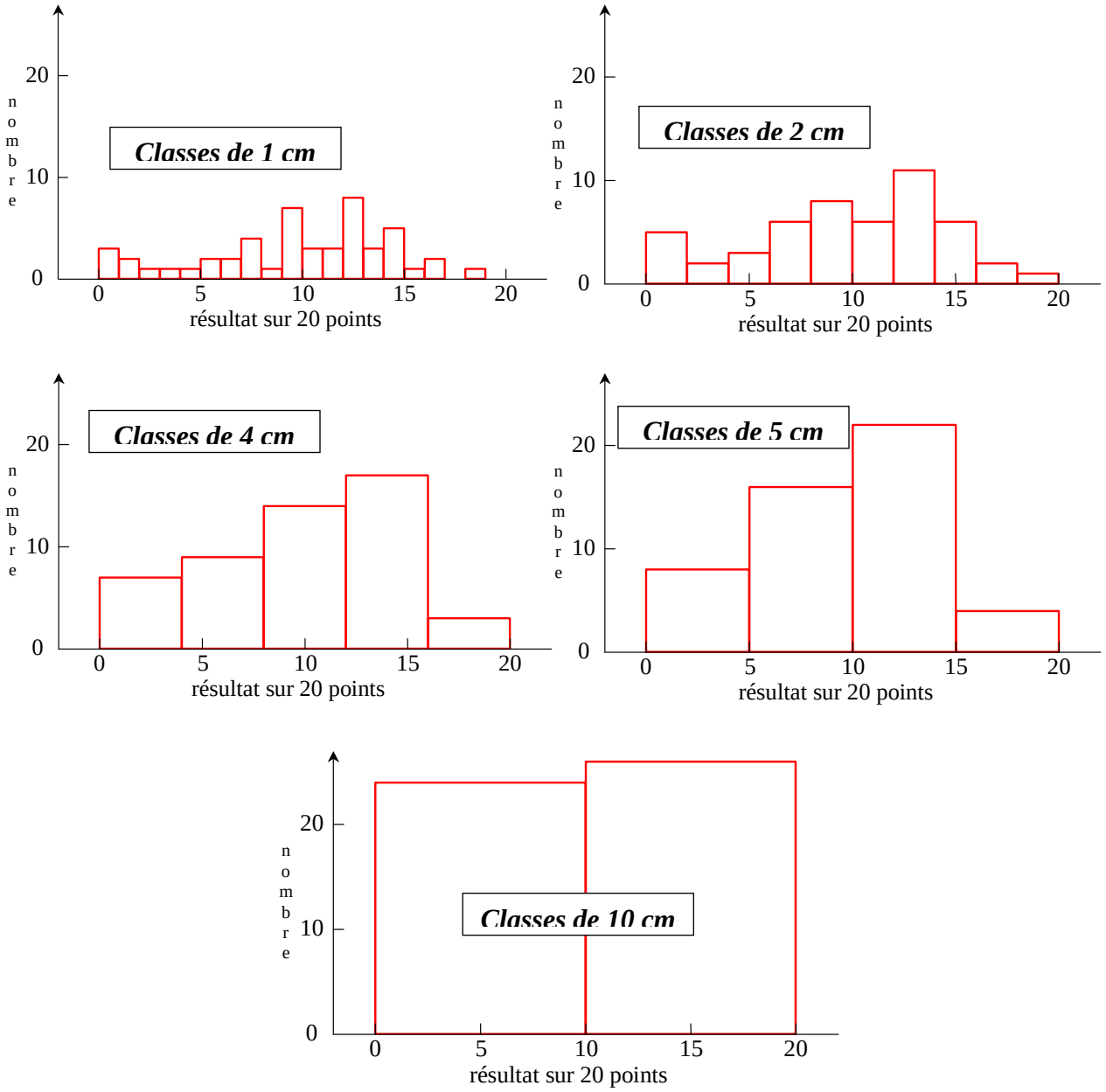
Ces considérations prennent leur sens dans le cadre de variables quantitatives.

Exemple :

Les cotes obtenues à un examen par 50 élèves sont données dans un tableau.

0.0	2.1	6.1	7.8	9.5	10.4	12.1	12.8	13.9	14.8
0.0	3.2	6.2	8.2	9.6	10.5	12.4	12.8	14.2	15.5
0.5	4.5	7.2	9.1	9.9	11.1	12.5	12.9	14.6	16.1
1.2	5.3	7.2	9.1	9.9	11.8	12.6	13.0	14.7	16.8
1.7	5.3	7.4	9.5	10.1	11.9	12.6	13.7	14.7	18.2

L'allure de l'histogramme change en fonction de la largeur choisie des classes, conditionnant l'interprétation, au moins visuelle, des résultats. Si, à partir de ces histogrammes, on trace le polygone de fréquence de ces représentations, le débat présentement suggéré devient encore plus évident ; sans parler encore des éventuelles déformations…

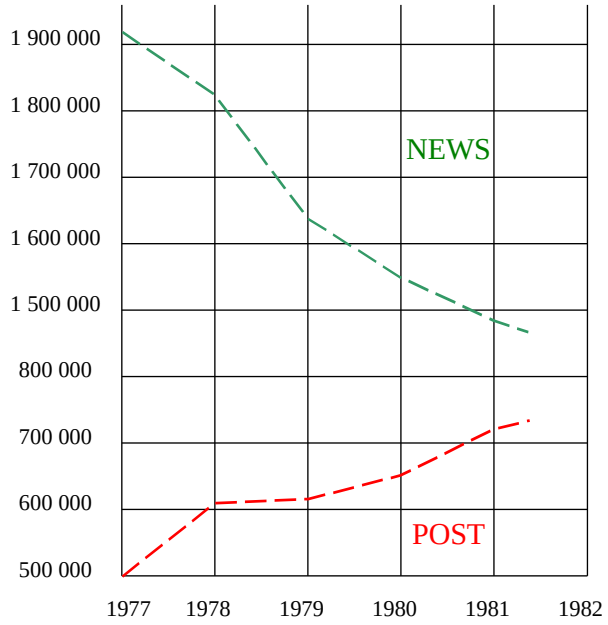


2. Bien interpréter les graphes.

Il est courant d'entendre déclarer que l'on fait dire aux statistiques ce que l'on veut. Par exemple, il est possible de présenter les résultats de manière à amener le lecteur peu attentif à accepter une conclusion erronée. Le but de ce chapitre est d'illustrer cette pratique par quelques exemples afin de donner des clefs pour interpréter correctement les graphes, parfois trompeurs.

- *Tirage de journaux concurrents :*

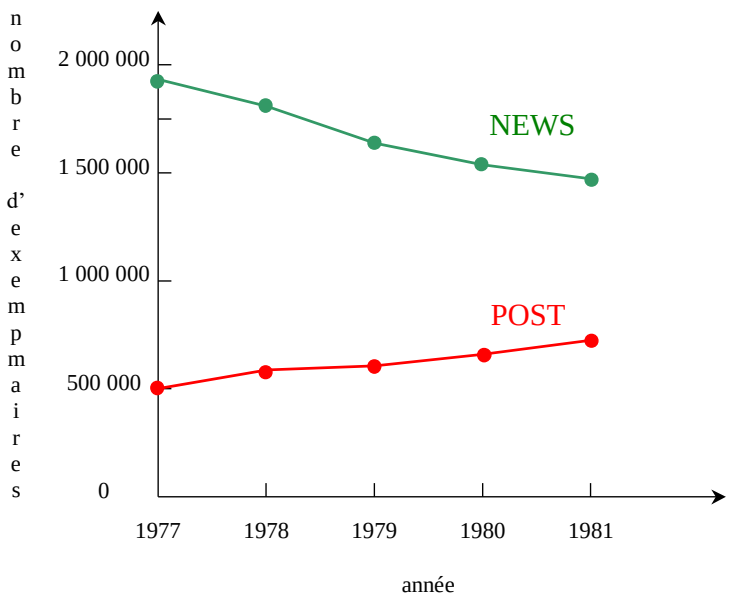
  Le graphique suivant est paru en 1981 dans le *New Yorker Post*, sous le titre « Ascension du Post, le quotidien préféré des New-Yorkais » et emprunté comme les suivants à feu mon professeur de Statistiques.



Le but de ce graphique est de convaincre le lecteur que la croissance du tirage du *Post* va bientôt l'amener en première position, devant le *News* qui périclite. On remarque deux artifices utilisés pour exagérer la tendance :

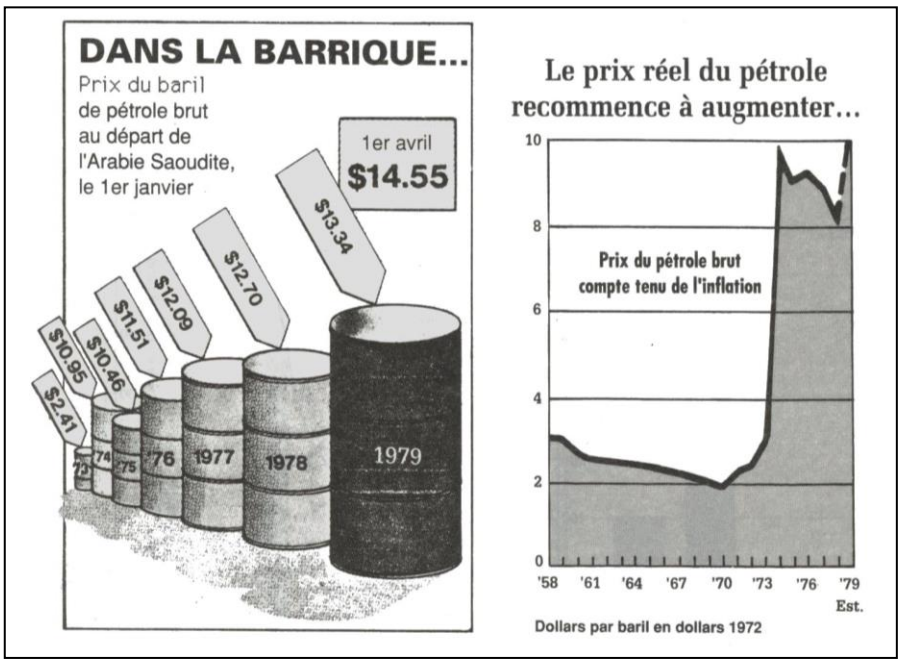
- l'échelle verticale ne démarre pas en zéro.
- l'échelle verticale est discontinue.

C'est une présentation acceptable, mais qui renforce les variations apparentes. Alors que deux graduations successives sont séparées de 100.000 unités, on passe brutalement de 800.000 à 1.500.000 dans l'intervalle séparant le *Post* du *News*. Les tirages des deux journaux paraissent, de ce fait, beaucoup plus proches que dans la réalité. Cette présentation ne serait admissible que si la discontinuité de l'échelle était clairement indiquée, par exemple par des pointillés. La version plus « honnête » du graphique montre qu'il reste au *Post* bien du chemin à parcourir avant d'accéder à la 1ère place.



- *Le baril de pétrole ... géant :*

  La figure ci-dessous, parue dans le Time du 9 avril 1979, est destinée à illustrer l'augmentation du prix du pétrole suite à la crise déclenchée par la guerre du Kippour.



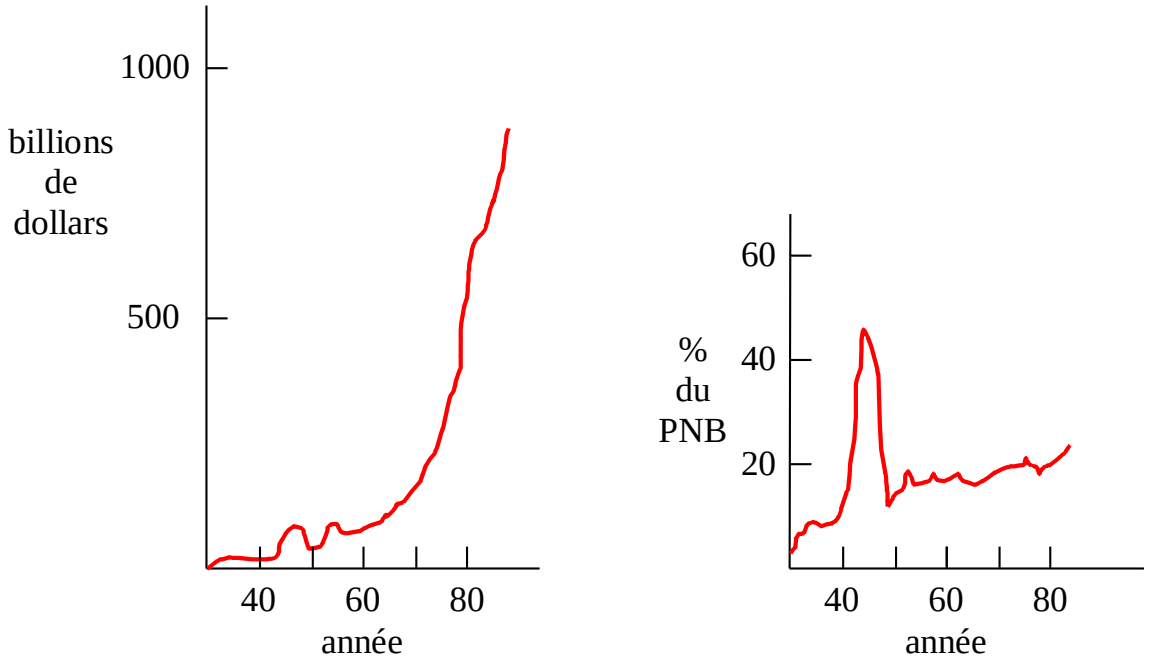
  De 1973 à 1979, le prix du pétrole a été multiplié par 6. Or, le baril « 1979 », qui est 6 fois plus haut que le baril « 1973 » contient 6 x 6 x 6 = 216 fois plus de pétrole que le premier. Ce n'est pas la hauteur du baril, mais son volume, que le lecteur associera généralement au prix. Le pétrole se vend au litre, pas au mètre ! On a donc exagéré d'un facteur 36 l'augmentation du prix du pétrole. Si, de plus, on tient compte de l'inflation, présentée dans la figure de droite, le prix du pétrole n'a augmenté que d'un facteur 3,5 entre 1973 et 1979. L'exagération est de 60 fois !

- *Dépenses gouvernementales aux Etats-Unis :*

  Le graphique suivant illustre l'accroissement des dépenses gouvernementales de 1930 à 1980. On constate une augmentation régulière si on mesure ces dépenses en dollars. Cependant, la mesure des dépenses en dollars n'a pas beaucoup de sens car elle ne tient pas compte de l'inflation et/ou de la croissance.

  Ce qui est plus significatif dans ce cas, c'est l'évolution des dépenses gouvernementales par rapport[39] à toutes les autres dépenses, mesurées ici, et par exemple, en pourcentage du Produit National Brut.

[39] Des liens, toujours des liens !



- *Production alimentaire mondiale :*

  Le graphe suivant, publié dans l'hebdomadaire *Business Week* en 1975, est destiné à illustrer la variation de la production alimentaire, comparée à celle de la population mondiale.

  La plupart des personnes examinant ce graphe vont conclure que la production alimentaire a augmenté moins vite que la population. Le piège réside dans le fait de comparer la production alimentaire *par tête*, soit par individu, à la population *totale*. Si la production alimentaire par tête augmente, cela signifie forcément que la production totale augmente plus vite que la population totale.

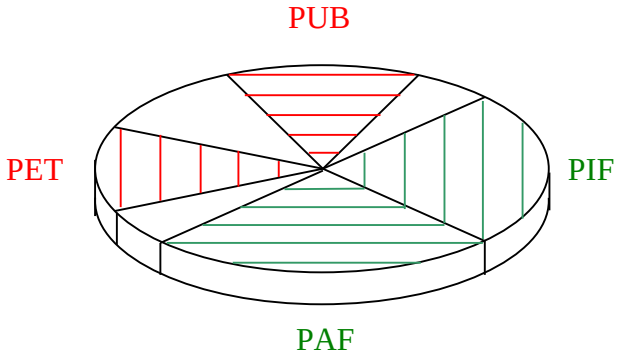
Une version plus claire de ce graphe est présentée ci-dessous :

NB : Il faut bien se garder d'interpréter les graphes au-delà de ce qu'ils présentent.

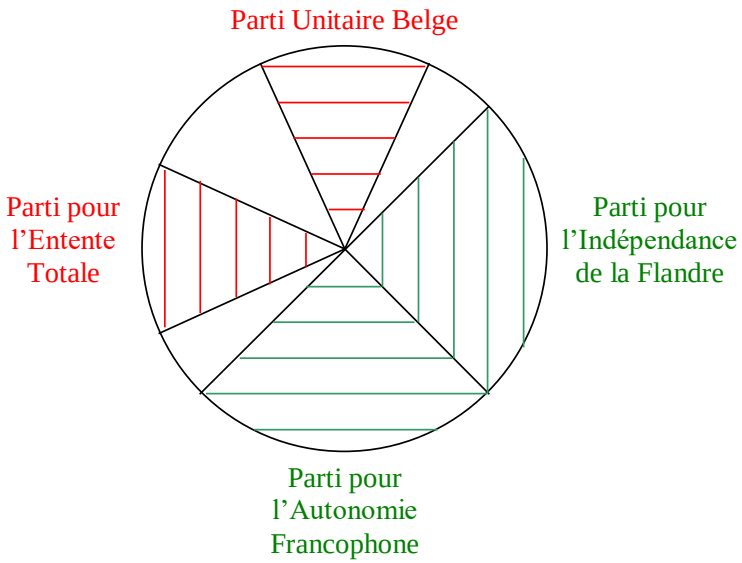
Du graphe ci-dessus, on ne peut pas déduire, par exemple, que le problème de la faim dans le monde était moins aigu en 1974 qu'en 1954. En effet, ce problème dépend de bien d'autres facteurs, comme la répartition des denrées alimentaires entre pays et entre couches de la population.

- *Le camembert en perspective :*

Le diagramme sectoriel suivant présente les pourcentages obtenus par quatre partis politiques lors d'une élection.

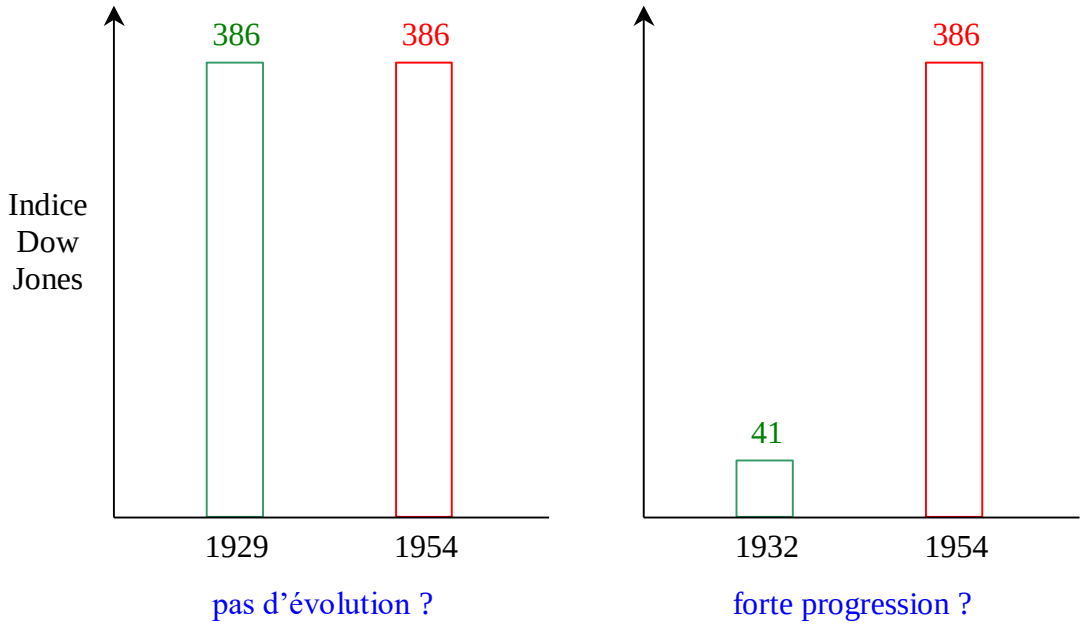


Une telle présentation en perspective a tendance à faire paraître plus importants les secteurs situés en bas (comme le PAF) ou en haut (comme le PUB) au détriment de ceux de gauche (PET) ou de droite (PIF). Une présentation de face est moins susceptible d'induire le lecteur en erreur. Cette remarque vaut aussi pour les histogrammes, surtout si ils sont cumulés.

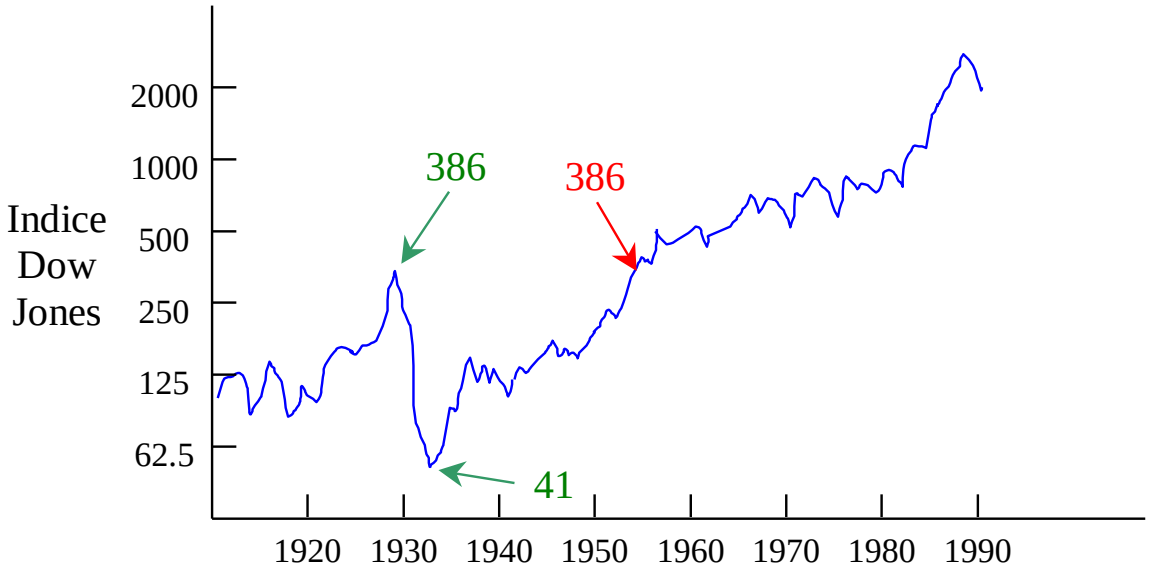


- *Choix de l'année de base :*

  L'évolution du marché boursier à *Wall Street* avant 1954 est illustrée sur deux graphiques, bien différents, ci-dessous :



  En regardant le graphe de gauche, on a l'impression que l'indice *Dow Jones* n'a pas évolué. Par contre, le graphe de droite suggère une forte progression. Ces deux graphiques, trop schématiques, donnent une vue tronquée de l'évolution du marché boursier. En examinant l'évolution complète de celui-ci, on constate que les années 1929 et 1932 prises comme références correspondent en fait à un pic et un creux de la courbe, la grande crise de 1930 ayant provoqué l'effondrement du cours des actions.

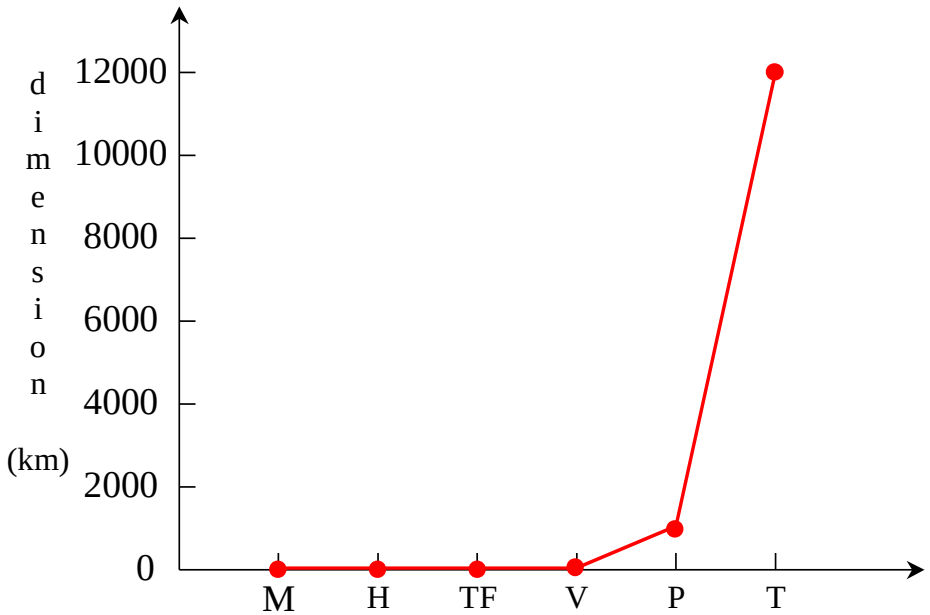


- *Echelle logarithmique :*

  Lorsque la grandeur à représenter varie fortement[40], l'échelle habituelle, linéaire n'est pas bien adaptée à la représentation des petites mesures de la variable considérée.

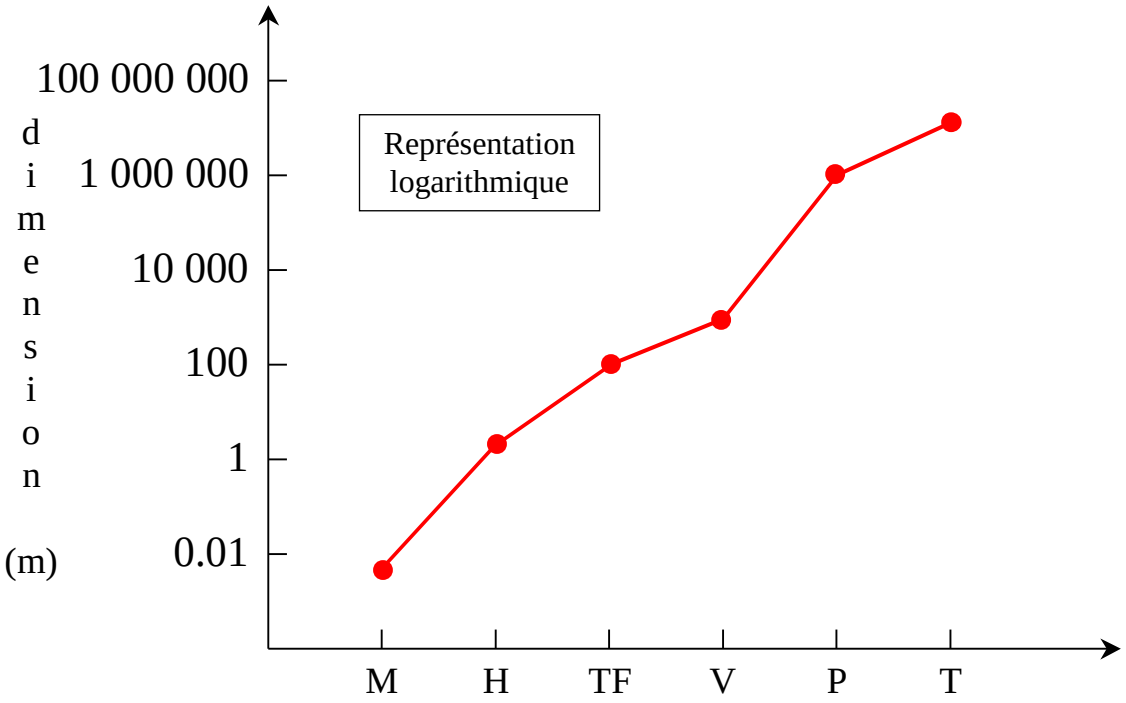
  Exemple : les dimensions des objets suivants sont : mouche (5 mm), homme (2 m), terrain de foot (100 m), village (1 km), pays (1.000 km), Terre (12.000 km).

  Dans une représentation linéaire, où une longueur donnée, entre deux graduations successives, correspond à l'addition d'une quantité fixée, les petites variations sont indiscernables. Ainsi, le graphique ci-dessus ne permet pas de distinguer la dimension d'une mouche de celle d'un terrain de football.



[40] de plus d'un facteur 100, par exemple.

Dans la représentation logarithmique, une distance fixe, entre deux graduations successives, correspond à la multiplication par un nombre donné[41]. Cette représentation est mieux adaptée à la comparaison des valeurs relatives.



[41] Ici, 100.

Chapitre 12 : Notion de loi statistique.

Après avoir étudié les différents indicateurs qui visent à décrire, avant de représenter, une distribution, approchons maintenant la notion de loi statistique. Il existe un certain nombre de lois statistiques qui permettent d'effectuer un certain nombre de calculs. Nous nous attarderons sur trois d'entre elles[42] : la loi de Gauss dite « normale », la loi de *Student* et la loi du Chi carré. Ces lois permettent de réaliser des tests statistiques dits d'inférence. Il s'agit de tests de comparaison, de tests d'hypothèse, de tests paramétriques. Il s'agit des appelations différentes d'une même pratique. Ces mêmes lois serviront également à l'estimation de la moyenne vraie.

1. L'inférence statistique :

   Il existe différences formes d'inférence ; inférer, c'est tirer une conclusion en mettant une proposition en relation. En statistique, cette inférence se veut démonstrative et est donc déductive[43]. Elle se porte, comme la validité que l'opération confère, sur deux niveaux, cités dans l'ordre d'inférence : interne puis externe.

   Ainsi, lorsque vous souhaitez comparer statistiquement deux groupes de sujets par rapport à leur positionnement sur une variable ou un caractère, lorsque vous recherchez le type de relation qu'entretiennent deux caractères d'une même variable, il faut choisir un test de comparaison. De même, lorsque vous souhaitez vérifier la validité de votre hypothèse expérimentale comme de chacune de vos hypothèses opérationnelles, il s'agit également de choisir un test d'hypothèse.

   Le principe de ce test est de pouvoir rejeter l'hypothèse nulle ou, autrement dit, de mesurer l'effet du hasard. En effet, la causalité ne peut être imputable au hasard et le rejet de l'hypothèse nulle est souvent un exercice primordial de l'outil statistique : validité interne.

   Lorsque vous souhaitez généraliser, inférer vos résultats, il convient de choisir une loi de référence afin de mesurer l'écart entre la moyenne mesurée, celle de l'échantillon, et la moyenne vraie, celle de la population. La validité externe de votre enquête est, alors et ensuite, interroger par ces tests où il s'agit de tester la probabilité d'effet du hasard ou, mieux, de la variable causale étudiée.

   Les probabilités sont régies par une série de lois : loi normale ou loi de Gauss, loi de *Student*, loi du Chi carré, … Les indices centraux et de dispersion qui permettent de caractériser une distribution vont, à nouveau, être utiles et seront retrouvés dans ces lois. Ces tests peuvent être réalisés, comme les précédents, par des logiciels mais quelques notions s'imposent pour choisir

[42] Le coefficient de Spearman (corrélation) fait également référence à une loi de ce type.

[43] L'inférence peut être inductive ou encore abductive. La logique comme la rhétorique utilisent ce mode de raisonnement mais à l'aide, cette fois, de notions et de concepts.

et en comprendre le résultat. Le choix est principalement déterminé par le type de variables et la taille de l'échantillon.

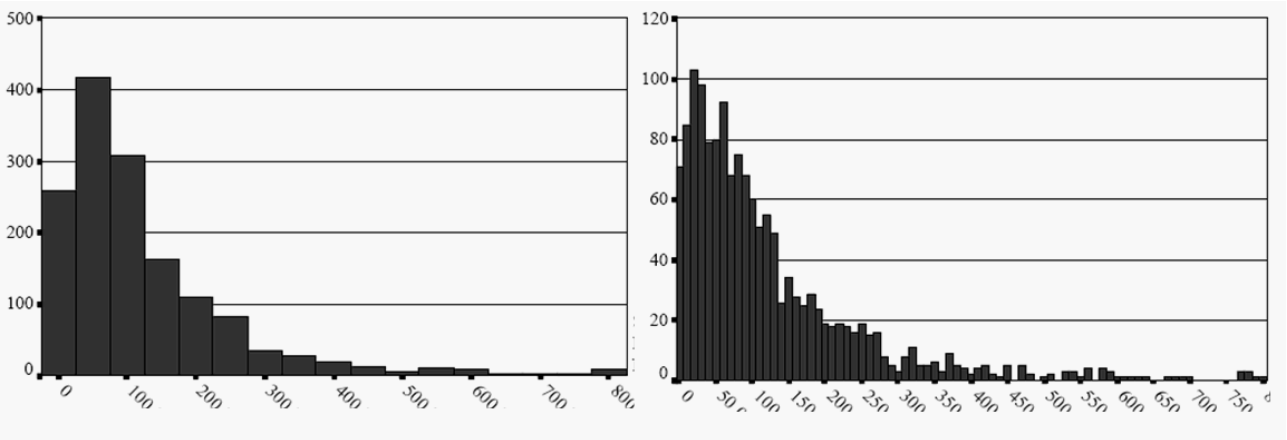
Remarque : il ne s'agit, en aucun cas, de confondre analyse et même inférence statistique avec interprétation des résultats[44].

2. La loi de Gauss, la « reine-mère » :
   - Bref rappel :

     Un des buts de la statistique est de construire un tableau de répartition, qui donnera l'effectif des individus entre les différentes valeurs possibles de la variable. Si la variable est discrète, le nombre de possibilités est limité. Si la variable est continue, le nombre de possibilités est illimité ou infini, même si elle est bornée. L'objectif de cette opération est d'obtenir la distribution de cette variable au sein de l'échantillon et/ou de la population. Cette distribution peut conduire à une représentation graphique.

     Comme suggéré dans le point concernant les quantiles, il est rare que les individus se distribuent également entre les différentes modalités ou mesures de la variable. Ce tableau de répartition peut encore bénéficier de classes par l'utilisation d'intervalles. Le découpage en classe est, généralement, régulier. Un découpage plus fin ou plus large est parfois choisi aux extrémités de cette distribution. Rappelons que les calculs s'opèrent alors sur le centre de la classe.
   - Distribution et répartition « normales » :

     Prenons ensuite la représentation graphique d'une distribution d'une variable quantitative, c'est-à-dire un histogramme. Comme vu précédemment, nous serons sensibles au fait que la représentation graphique et l'analyse (et …) dépendent étroitement du niveau de découpage de la série. Un autre exemple est utilisé ci-dessous où les intervalles de la seconde figure sont cinq fois plus petits.



[44] Dans le cas contraire, c'est à nouveau confondre l'outil et l'artisan !

La distribution est ici très asymétrique, comme le soulignait déjà la loi de Pareto[45]. Il s'agit d'un exemple économique. Cette asymétrie est un phénomène courant et conduit à un type d'analyse : l'analyse de concentration et son versus, l'analyse de dispersion avec leurs indices de résumé respectifs.

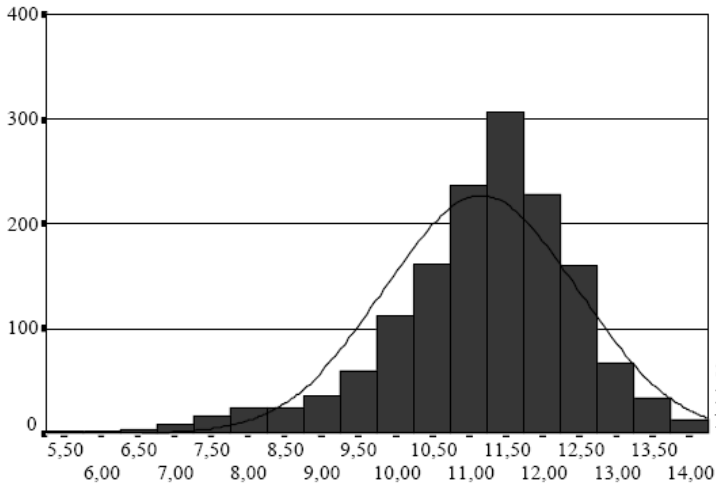
L'explication de ce phénomène, récurrent, a conditionné de nombreuses pratiques statistiques et de nombreuses pratiques (tout court). Elle tient dans « une » observation faite dès le XVIII^{ème} siècle par Gauss et Laplace, simultanément et indépendamment ; et formulée de la manière suivante : « *si un phénomène résulte d'une quantité de causes infinie dont les effets sont comparables et additifs, les valeurs de la variable prises par une population d'individus se distribue en suivant la loi normale, dont la forme caractéristique est celle de la cloche ; par contre si ces causes sont de nature multiplicative,* (au lieu de s'ajouter, elles se multiplient) *alors on observe une asymétrie de distribution* ».

Que comprendre de cette assertion ?

Sans entrer dans le détail des règles mathématiques, la fonction logarithmique d'un nombre s'additionne Si Yi = X1 * X2 * X3 * … , alors log(Yi) = log(X1) + log(X2) + log(X3) + …

> Remarque : il est prudent de ne pas confondre les différents types de logarithmes : logarithme népérien (fonction notée « ln » y compris dans Excel®) et le logarithme décimal (log10) ou autrement dit le logarithme[46] à base 10.

Dès lors, on peut s'attendre à ce que l'histogramme du logarithme de la valeur observée retrouve la forme caractéristique de la loi de Gauss et redevienne symétrique : la fameuse « *curve bell* ». Cette transformation logarithmique est donc souvent nécessaire puisqu'un grand nombre de modèles statistiques requiert une distribution « normale ».



[45] W. Pareto, économiste connu et reconnu pour la notion d'optimum, il a sa place dans notre cours d'*Economie politique de la Santé* et une, plus modeste, dans *Sociologie de la Santé*.

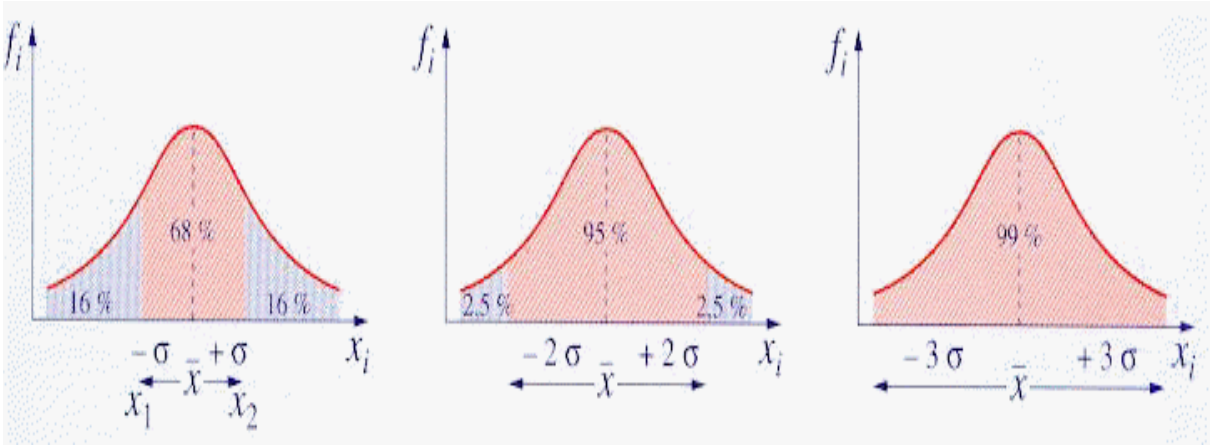
[46] Nous avions choisi log100 dans un chapitre précédent.

- Gauss et normalité d'une moyenne :

  Comme la plupart des lois utilisées en statistique, la loi normale est une loi de probabilité. Cette loi normale est souvent qualifiée de loi du hasard[47]. Autrement dit, c'est le hasard modélisé, géré et assuré par la loi de Laplace-Gauss. Cette loi a influencé de nombreux autres modèles ; non seulement au sein de la statistique mais également dans les autres disciplines qui en usent : économie, sociologie, politique. En effet, du hasard, elle modélise des notions comme celles du risque, des aléas, de l'intervalle de confiance, etc..

  La courbe de Gauss est, en réalité, une distribution théorique, considérée dès lors comme normale, c'est-à-dire reprenant tous les cas. La moyenne, autre notion « normalisante » est au centre de cette représentation graphique séparant d'abord la majorité, les minorités et les extrêmes en fonction de l'écart type.

  Comme nous l'avons vu précédemment, cette distribution est caractérisée par deux éléments : un indice central, la moyenne (m) et un indice de dispersion ou d'étalement, l'écart type (σ). La répartition de cette distribution se fait symétriquement à la moyenne médiane en fonction de l'écart type.



  L'application, réitérée, de celle-ci conduit à l'élaboration d'une table[48] qui permet de calculer facilement cette fonction. Pour être précis sur le plan terminologique, recourir à cette table s'est solliciter la « loi normale centrée ». Cette distribution théorique, ce modèle, cette référence utilise donc cette table et la répartition se fait de part et d'autre de la moyenne en fonction de sa carte de contrôle :

[σ : 68 %], [2σ : 95 %], [3σ : 99 %]
ou
[σ : 64 %], [2σ : 96 %], [3σ : 99,8 %]

[47] ou « loi de probabilité de la variable aléatoire » ou encore « théorème de la limite centrale ».
[48] Voir en fin de balise.

Ainsi, une des caractéristiques de cette loi dite normale est de trouver aux extrémités de cette distribution de très grandes valeurs. Ces valeurs extrêmes ont un point commun : leur fréquence de survenance est aussi faible que leur magnitude est forte. Ces événements n'arrivent pas souvent, mais ils coûtent cher ou rapportent beaucoup. Ici et ainsi, Gauss et Pareto se rejoignent à nouveau.

3. Relativité de normalité

Nous l'avons souligné, cette loi a influencé de nombreux autres modèles, en particulier en économie. Même si les modèles usuels de l'économie sont des modèles dans lesquels les événements rares n'existent pas. Seuls les événements moyens sont pris en compte et analysés, et cette « moyennisation » forcée permet un traitement rassurant des risques liés au hasard : les aléas. De façon générale, les modèles classiques de l'économie[49] sont tous fondés sur cette notion de moyenne, faisant ainsi passer la notion d'espérance mathématique à celle de probabilité. Nous vivons ainsi dans un monde réel dont le risque est géré par des probabilités. Rassurant, le modèle, donc assurables, les risques. Le modèle-type d'application de cette logique sont les assurances.

Or, le monde réel ignore les moyennes, qui sont des inventions[50] de l'esprit humain comme les mathématiques. Ainsi, ces considérations font apparaître extrêmement clairement que le calcul d'une moyenne n'est pertinent que si la dispersion autour de la moyenne est limitée.

4. Retour sur la statistique :

Et les joies du calcul ! Les lois comme celle de Gauss est l'extension à l'infini d'un calcul de probabilité que nous appliquerons au jeu de dé. Dans le lancer d'un dé, la variable aléatoire ne peut prendre que six valeurs ; il s'agit donc d'une variable discrète. Si le dé n'est pas truqué, la loi de probabilité se résume :

x_i	1	2	3	4	5	6
$p(x_{i)}$	1/6	1/6	1/6	1/6	1/6	1/6

Dans la plupart des expériences, on a rarement accès à l'expression exacte de la probabilité pour plusieurs raisons dont la principale tient au fait de l'échantillonnage. On se contente donc de calculer un certain nombre d'indicateurs, dont le plus caractéristique est la moyenne. Tous ces indicateurs peuvent être obtenus de deux façons différentes :

[49] D'autres disciplines, sciences pourtant humaines (*sic*), tendent à y recourir par leurs pratiques hérités de l'expérimentalisme, réclamant la preuve par les chiffres. Depuis Claude Bernard, la médecine est largement influencée, pour ne pas dire déterminée, par le paradigme quantitativiste.

– Si la loi de probabilité est connue *a priori*, alors on peut les calculer directement à partir de cette loi. On obtient alors une valeur exacte ou théorique. C'est la cas du dé ci-dessus.
– Si la loi de probabilité n'est pas connue, alors il faut réaliser « une » expérience pour estimer les indicateurs. Les valeurs obtenues seront d'autant plus proches des valeurs théoriques que l'expérience a été bien menée. Nos recherches se situent toujours dans ce cas.

Pour les puristes, si la loi de probabilité est connue, la moyenne théorique se nomme espérance (μ). En reprenant l'exemple du dé non truqué, on obtient puisque la loi de probabilité est connue :

x_i	1	2	3	4	5	6
$p(x_i)$	1/6	1/6	1/6	1/6	1/6	1/6
μ = (1 * (1/6)) + (2 * (1/6)) + (3 * (1/6)) + (4 * (1/6)) + (5 * (1/6)) + (6 * (1/6)) = 7/2						

L'espérance est de 3,5. Ce résultat est exact et ne dépend pas du nombre de lancers. En réalisant l'expérience pour des nombres de lancers différents, on peut obtenir, par exemple :

n	10	100	1000	10000	100000
m	3,2	3,56	3,486	3,5366	3,5010

Ces valeurs convergent vers le résultat théorique pour n → ∞. Une précision se répète : il existe d'autres lois de probabilité appelées aussi aléa mais Gauss et *Student*, la « petite sœur » de la « reine mère », restent les plus communes. Ces deux lois, outre leur différence, se rencontrent fréquemment et s'appliquent à tous les phénomènes qui résultent d'un grand nombre d'évènements indépendants et d'origines diverses. L'explication se trouve dans le théorème de la limite centrale.

Soit X une variable aléatoire de moyenne μ, de variance σ^2 et dont la loi de probabilité est quelconque. Soit

$$y_N = \frac{1}{N} \sum_{i=1}^{N} x_i$$

une moyenne effectuée sur un grand nombre N de mesures. Si σ^2 est fini, alors la distribution de y_N tend vers une loi normale de moyenne μ et de variance σ^2/N.

[50] Des construits, disait-on en sociologie !

Ce théorème peut s'interpréter comme suit : si une grandeur subit l'influence d'un nombre important de facteurs indépendants, et si l'influence de chaque facteur pris séparément est petite, alors la distribution de cette grandeur tend vers la loi normale. Les deux premiers éléments de cette assertion définissent en quelque sorte le hasard.

C'est ainsi qu'une fois le hasard modélisé, il devient possible de « calculer » sa présence ou son absence ; comme le degré de certitude d'une valeur, appelée intervalle de confiance. Cet intervalle de confiance suggère la marge d'erreurs, au moins de mesure entre l'échantillon et sa population d'origine. La valeur de cette incertitude est toujours approximative ; on se contente de la représenter en équilibrant les décimales. Il s'agit ici d'une convention et d'une règle d'homogénéité.

Exemple : Pour un résultat calculé, on obtient : 9,8188 ± 0,032554 ; on écrira : 9,81 ± 0,03.

Pourtant, cet exercice de détermination ne mesure pas toutes les marges d'erreurs ; et donc toutes les erreurs possibles. Ces tests paramétriques ne peuvent estimer, en particulier, les erreurs liées au dispositif de recherche[51] ou si elles se propagent[52]. Notons qu'il existe également certains modèles qui permettent de l'évaluer mais aucun d'entre d'eux ne peut prétendre mesurer l'incidence des biais.

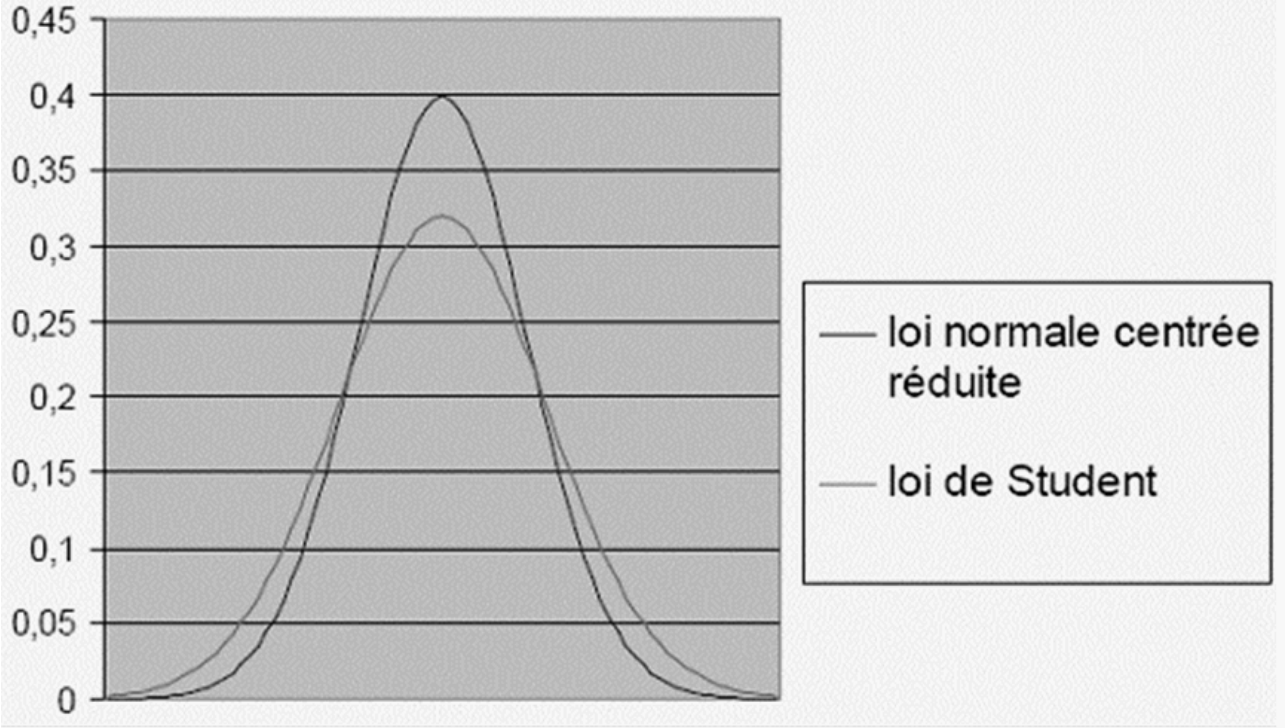
La loi normale et les autres dérivées[53] permettent d'opérer des tests statistiques de deux ordres :

- les tests paramétriques d'hypothèse, correspondant au *temps 1* de l'inférence et découlant sur la validité interne.
- l'estimation ou calcul de l'intervalle de confiance, correspondant au *temps 2* de l'inférence et découlant sur la validité externe.

Nous ne pouvons rater l'occasion de dire et répéter que ces étapes de validation permettent, autorisent l'interprétation des résultats, puis de leur généralisation.

Avant de poursuivre, soulignons d'abord que le but des tests d'hypothèse n'est pas de choisir entre l'hypothèse nulle (H0) et l'hypothèse alternative, c'est-à-dire celle qui est expérimentée, mais plutôt de rejeter H0, celle qui résulte du hasard. Reprenons enfin cette distinction fondamentale entre loi de *Student* et loi normale. Nous connaissons la courbe de Gauss et ces critères liés à l'écart-type. La courbe de la loi de *Student* est très proche de celle de la loi normale. Seulement, l'écart-type est un peu plus grand que celui de la loi normale ; la courbe s'en retrouve de ce fait plus aplatie.

[51] Exemple : la montre utilisée retarde d'une seconde par minute.
[52] Biais de propagation : l'enchaînement d'erreurs se potentialisent !
[53] *Student, Bravais Pearson* et Chi carré, pour notre petite part.



La loi de *Student* tend vers la loi normale ; autrement dit, la loi normale prend le relais lorsque la taille de l'échantillon dépasse **30**. La loi normale nous permet donc de tester les hypothèses dans le cas d'un échantillon d'un effectif supérieur à 30 et d'agir de même pour l'estimation de l'intervalle de confiance. Sous cette barre, nous aurons recours à la loi de *Student*.

Le type de variables statistiques conditionne le choix des tests et calculs à utiliser, selon qu'elles soient quantitatives, alors nous travaillerons des moyennes arithmétiques ; ou qu'elles soient qualitatives, il s'agira de manipuler des proportions. Bien que pour ces dernières, il est possible de se référer aux mêmes principes que pour les variables quantitatives, il reste toutefois préférable d'utiliser le test du Chi carré considéré comme plus congruent.

Chapitre 13 : Tests paramétriques ou tests d'hypothèse

Dans la suite des lois statistiques, nous approcherons maintenant les tests paramétriques qui se distinguent selon le type de variables et la taille de l'échantillon. Sur ce point, nous avons lourdement insisté. Chacun d'eux, conditionnés par les deux éléments susmentionnés, sollicite une loi de probabilité. Notre usage de la panoplie sera limité et donc non exhaustif.

Le principe de tout test de ce type est de confronter deux hypothèses afin d'en invalider l'une, validant ainsi la seconde. Il convient donc, après avoir fait le choix du test, de calculer le paramètre du test. Est associé au paramètre du test un risque grâce à une table de distribution du paramètre[54]. Il permet de prendre une « décision » à partir d'informations fournies par un échantillon.

L'hypothèse considérée comme nulle est celle qui s'éloigne, voire s'oppose à l'expérimentation et à la réponse qu'elle suppose et attend. L'hypothèse nulle est généralement une hypothèse « conservative » ; à moins qu'il ne s'agisse du hasard.

Exemple : *un nouveau médicament sera jugé efficace s'il apporte clairement une amélioration. On testera donc H_0 : le médicament est inefficace contre H_1 : le médicament est efficace.*

Une procédure de test engendre deux types d'erreurs :

- l'erreur de première espèce (α) conserve H_0 alors qu'elle est fausse. C'est le faux négatif où est en cause la sensibilité de l'expérimentation ;
- l'erreur de seconde espèce (β) rejette H_0 alors qu'elle est vraie. La spécificité de l'expérimentation conduit à la production de faux positifs.

Face à ces erreurs, on souhaite minimiser le risque d'erreurs, en particulier les α.

Le but du test paramétrique n'est pas de choisir entre les deux mais de pouvoir rejeter H_0.

<table>
<tr><td colspan="2">Variables QUANTITATIVES</td><td>Variables QUALITATIVES (*)</td></tr>
<tr><td colspan="2">Tests de comparaison de moyennes</td><td>Tests de comparaison de proportions</td></tr>
<tr><td colspan="2">Taille de l'échantillon</td><td>⇨ le paramètre du test suit :</td></tr>
<tr><td>Grands échantillons
(n>30)</td><td>Petits échantillons
(n<30)</td><td rowspan="3">la loi normale réduite
à condition que :
n p ≥ 5 et n (1 - p) ≥ 5.</td></tr>
<tr><td colspan="2">⇨ le paramètre du test suit :</td></tr>
<tr><td>la loi normale</td><td>la loi de Student</td></tr>
</table>

(*) Sur des variables binaires, à deux modalités où « 1 » pour l'événement, « 0 » pour son absence, on s'intéresse à l'une des deux proportions p (1 – p = q). Au-delà, la technique devient plus ardue.

[54] de sa probabilité.

Plusieurs tests sont envisageables dans chacune de ces trois catégories. A ceux-ci, il faut ajouter le test statistique propre aux corrélations et un autre, plus spécifique aux variables qualitatives :

	Test de comparaison	**Situation**
Variables quantitatives + Grands échantillons = loi **normale**	d'une moyenne observée à une valeur donnée	1 échantillon ➔ 1 moyenne observée
	de deux moyennes observées	2 groupes distincts ➔ 2 échantillons ➔ 2 moyennes observées
	de moyennes sur séries appariées	1 groupe ➔ 2 échantillons[55] ➔ 2 moyennes observées
Variables quantitatives + Petits échantillons = loi de **Student**	d'une moyenne observée à une valeur donnée	1 échantillon ➔ 1 moyenne observée
	de deux moyennes observées	2 groupes distincts ➔ 2 échantillons ➔ 2 moyennes
	de moyennes sur séries appariées	1 groupe ➔ 2 échantillons[22] ➔ 2 moyennes observées
Variables qualitatives = loi **normale réduite**	d'une proportion observée à une valeur donnée	1 échantillon ➔ 1 proportion
	de deux proportions observées	2 groupes ➔ 2 échantillons de données ➔ 2 proportions
	de proportions sur séries appariées	1 groupe ➔ 2 échantillons[22] ➔ 2 proportions observées

1. <u>Comparaison d'une moyenne observée à une moyenne théorique (grands échantillons)</u> :
   Hypothèse nulle : l'échantillon a été prélevé dans une population dont la moyenne est égale à la moyenne théorique et la différence observée est due au hasard de l'échantillonnage.

$$\epsilon = \frac{\bar{X} - X_0}{s/\sqrt{n}}$$

[55] Un seul et même groupe sur lequel on effectue deux séries de mesure.

2. <u>Comparaison de deux moyennes observées (grands échantillons)</u> :

Exemple : on considère que deux classes qui ont les caractéristiques suivantes :

- classe A de 34 élèves : moyenne semestrielle 10,2 (s = 3,5),
- classe B de 32 élèves : moyenne semestrielle 12,4 (s = 3).

⇨ Peut-on dire que ces deux classes sont d'un niveau significativement différent ?

⇨ Les deux échantillons ne proviennent-ils pas de la même population ?

	Echantillon A	Echantillon B
Moyenne (X)	10,2	12,4
Ecart type (σ)	3,5	3
Effectif (n)	34	32

Hypothèse nulle : les deux échantillons proviennent de la même population et les différences observées sont dues au hasard de l'échantillonnage.

Les effectifs sont supérieurs à 30.

Le calcul de l'écart réduit

$$\epsilon = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{s^2{}_A}{n_A} + \frac{s^2{}_B}{n_B}}}$$

ε = ...

L'écart réduit vaut donc ..... et ..... la valeur limite de 1,96 correspondant à un risque de 5 %[56]. Sur le table, cette valeur correspond à un risque de ..... .

Conclusion : .....

3. <u>Comparaison de deux moyennes sur séries appariées (grands échantillons)</u> :

On se trouve dans le cas de séries appariées lorsque l'on effectue plusieurs fois la même mesure sur les mêmes individus. Les éléments à considérer pour ce test sont :

- $\overline{D}$, la moyenne des différences : $\overline{D} = (\Sigma (X^2 - X^1))/n = (\Sigma Di)/n$
- Vx_D, la variance de la moyenne des différences : $VxD = \Sigma ((Di - \overline{D})^2)/n$
- S ou σ est la racine carrée du calcul précédent : $\sqrt{Vx_D}$, soit son écart type.

L'écart réduit se calcule alors :

$$\epsilon = \frac{\bar{D}}{s/\sqrt{n}}$$

[56] Voir fin de balise : table de la loi normale centrée réduite simplifiée.

4. Comparaison d'une moyenne observée à une moyenne théorique (petits échantillons) :

Pour tout échantillon inférieur à 30 d'effectif, il faut recourir à la table du t de *Student* qui tient compte du nombre[57] « effectif » d'individus au sein de l'échantillon soumis à l'inférence. L'exercice vise ici à comparer une moyenne observée à une valeur théorique en calculant l'équivalent de l'écart réduit rapporté à la loi de *Student*(*-Fisher*) : le *t*. Le principe de fonctionnement reste identique la valeur obtenue sera confrontée à la valeur affichée dans la table de *Student*, dans ce cas précis, à un degré de liberté :

ddl = (n-1).

$$t = \frac{\bar{X} - X_0}{s_X / \sqrt{n}}$$

avec X = la moyenne observée,

X_0 = la moyenne théorique

avec S ou σ = l'écart type

n = l'échantillon d'un effectif inférieur ou égal à 30.

Exemple :

Les notes d'une classe de 10 élèves se répartissent comme suit : 12/8/9/7/13/6/8/11/9/10 et nous possédons une valeur de référence de 10/20. L'hypothèse nulle peut être formulée ainsi : *les notes issues d'une population dont la moyenne est de 10 sur 20 et la différence observée au niveau de notre échantillon est due au hasard de l'échantillonnage.*

m = ...

Vx = ...

t = ...

pour un ddl = n – 1 = ... avec un risque choisi à 5 %, la table indique : .....

Notre valeur est ..... à la valeur seuil de la table.

Donc, l'hypothèse nulle ..... rejetée.

Conclusion : ..... .

5. Comparaison de deux moyennes observées (petits échantillons) :

Ici, le degré de liberté sera calculé à $(n_A + n_B) - 2$

Ce test permet de comparer deux moyennes observées ; entre celle du groupe expérimental et celle du groupe-témoin.

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{s_G^2}{n_A} + \frac{s_G^2}{n_B}}}$$

[57] Pour les échantillons de taille supérieure à cette « limite », l'écart réduit suit la distribution de probabilité de la loi normale quel que soit le nombre d'individus qui composent l'échantillon.

Un exemple éclaire la formule. Comparons deux groupes de répondants à une échelle de satisfaction, le premier groupe est le groupe expérimental et le second, groupe témoin.

Valeurs	Effectif G1 : **A**	Effectif G2 : **B**
0	0	0
1	0	1
2	1	1
3	1	2
4	3	4
5	1	5
6	4	3
7	3	2
8	1	1
9	4	0
10	2	1
Total	$n_A = 20$	$n_B = 20$

Calcul décomposé du *t* de *Student* :

1°- Calcul des moyennes pour chaque groupe :

m_{G1} ou X_A = ... et m_{G2} ou X_B = ...

2°- Calcul des variances de chaque groupe :

G1 = échantillon A				G2 = échantillon B			
xi	xi - mi	$(xi - mi)^2$	$n(xi - mi)^2$	xj	xj - mj	$(xj - mj)^2$	$n(xj - mj)^2$
1				1			
2				2			
3				3			
4				4			
5				5			
6				6			
7				7			
8				8			
9				9			
10				10			
		$\sum_i n_i (x_i - m_i)^2$				$\sum_j n_j (x_j - m_j)^2$	

- variance pour G1 = Vx_A = ... & écart-type G1 = ...
- variance pour G2 = Vx_B = ... & écart-type G2 = ...

3°- Calcul de la variance globale (ou commune) :

$$= \frac{\sum_i n_i (x_i - m_i)^2 + \sum_j n_j (x_j - m_j)^2}{(N1 + N2) - 2} = ...$$

4°- Calcul du t de *Student* :

$$t = \frac{m1 - m2}{\sqrt{S_G^2 var\,(1/N1 + 1/N2)}} = ...$$

5°- Comparaison de la valeur du t au seuil de la table :

Comparer la valeur du t (.....) au seuil de la table pour un risque p fixé et en fonction du degré de liberté. Pour ..... ddl, le t lu dans la table indique : ..... pour p = 0,05 ; ..... pour p = 0,02 et ..... pour p = 0,01. Ainsi, le t de *Student* calculé (2,87) est ..... à celui lu dans la table pour un risque p = ...

6°- Interprétation : .....

6. <u>Comparaison de deux moyennes sur séries appariées (petits échantillons)</u> :

On se trouve dans le cas de séries appariées lorsque l'on effectue plusieurs fois la même mesure sur les mêmes individus. Les éléments à considérer pour ce test sont :

- $\overline{D}$, la moyenne des différences : $\overline{D} = (\Sigma (X^2 - X^1))/n = (\Sigma Di)/n$
- Vx_D, la variance de la moyenne des différences : $VxD = \Sigma ((Di - \overline{D})^2)/n$
- S ou σ est la racine carrée du calcul précédent : $\sqrt{Vx_D}$, soit son écart type.

Le *t* de *Student* se calcule alors :

$$t = \frac{\bar{D}}{s_D / \sqrt{n}}$$

<u>Exemple</u> : on mesure une variable avant et après l'introduction d'une procédure afin de vérifier l'influence de celle-ci sur cette variable.

Avant (pré)	Après (post)	Avant (pré)	Après (post)
52	53	48	49
34	43	48	53
56	59	56	59
43	46	55	55
43	46	43	42
52	53	56	61
51	55	46	52
45	49	56	64

Le calcul consiste à calculer d'abord les différences entre les premières mesures et les secondes pour chaque individu, puis la moyenne et la variance de cette différence.

Avant (pré)	Après (post)	D
	Moyenne	
	Variance	
	Ecart-type	

Le *t* de Student-Fisher vaut : *t* = ... = ... ; le ddl est ..... et la valeur limite dans la table est ..... . La valeur obtenue est ..... à la valeur tabulée, l'hypothèse nulle ..... être rejetée.

Conclusion : ..... .

7. Comparaison d'une proportion observée à une proportion théorique :

   Proportion observée : p

   Proportion théorique : p_0

   q = (1 - p)

$$\epsilon = \frac{p - p_0}{\sqrt{\frac{pq}{n}}}$$

8. Comparaison de deux proportions observées :

$$\epsilon = \frac{p_A - p_B}{\sqrt{\frac{pq}{n_A} + \frac{pq}{n_B}}}$$

avec

$$p = \frac{n_A p_A + n_B p_B}{n_A + n_B}$$

$$q = 1 - p$$

   nApA = effectif de l'une des valeurs de la variable binaires (oui, vrai,...)

   nBpB = effectif de la même valeur de la variable binaire (oui, vrai,...) dans l'autre échantillon.

   Rappel : p + q = 1

9. Comparaison de deux proportions observées sur séries appariées :

   - *a* est le nombre d'individus dont la réponse a évolué dans un sens ; (exemple : oui lors de la première mesure et non lors de la deuxième).
   - *b* est le nombre d'individus dont la réponse a évolué dans l'autre sens ; (exemple : non lors de la première mesure et oui lors de la deuxième).

$$\epsilon = \frac{a - b}{\sqrt{a + b}}$$

10. Comparaison et test paramétrique d'un calcul de corrélation :

    - Table de contrôle :

      Il s'agit de confronter le *r* calculé avec une table des valeurs de *r* servant de référence. La part de risque *p* est toujours identique, soit 5 %, on regarde si le coefficient calculé est supérieur ou égal au *r* lu dans la table pour un degré de liberté (ddl[58]) égal à n-2.

[58] Prudence, le degré de liberté peut être différent d'un test à l'autre.

Si le coefficient est inférieur à celui lu dans la table, alors il n'y a pas de corrélation significative entre les deux variables. Dans de nombreux cas, il peut s'agir d'une erreur d'échantillonnage : échantillon biaisé ou trop petit.

Table des valeurs significatives de *r* ou coefficient de Bravais-Pearson en fin de balise.

- Coefficient en rang de Spearman

En principe, le coefficient de Pearson n'est applicable que pour mesurer la relation entre deux caractères ayant une distribution dissymétrique et ne comportant pas de valeur exceptionnelle. Si ces conditions ne sont pas remplies, l'utilisation de ce coefficient n'est pas conseillée. Il est alors préférable d'utiliser le coefficient en rang de Spearman. Par ailleurs, l'absence d'une relation linéaire ne signifie pas l'absence de toute relation entre les deux caractères étudiés. Il est possible aussi de vérifier avec le coefficient de Spearman. Faut-il l'inscrire au titre des automatismes statistiques ?

Le coefficient en rang de Spearman ou *rho* de Spearman (noté « ρ ») est un coefficient de corrélation. Il est fondé sur l'étude de la différence de rangs entre les valeurs de deux ou plusieurs caractères étudiés. Il permet de rechercher l'existence de relations monotones, croissantes ou décroissantes, quelque que soit la forme du nuage de points. Il s'agit de remplacer chaque valeur d'un caractère par son rang dans la série observée et de calculer la différence de rang entre les deux mesures.

Comment procéder à ce classement ?

1°- Préciser l'ordre de classement : croissant ou décroissant. Il doit être le même pour les deux caractères.

2°- Lorsqu'il y a des ex æquo, le rang qui leur est attribué est celui de la moyenne des rangs qu'ils auraient occupés s'ils avaient été à la suite les uns des autres. On reprend ensuite le classement après les rangs qu'ils auraient pris s'ils n'avaient pas été ex æquo.

3°- La formule du coefficient *rho* est la suivante :

d représente la différence entre les rangs de 2 variables.

$$\rho = 1 - \frac{6 \Sigma d^2}{n(n^2 - 1)}$$

4°- Ce coefficient varie entre –1 et +1. Son interprétation est identique à celui de Pearson, mais il permet de mettre en évidence des relations non linéaires lorsqu'elles sont positives ou négatives.

Exercice :

Nous souhaitons rechercher s'il y a une corrélation entre la variable « *niveau de dépendance des personnes âgées de plus de 70 ans* » exprimé par les soignants sur une échelle de 0 à 10 et le « *niveau général d'anxiété de la personne* » exprimé par les patients sur une échelle de 0 à 10. La population observée est de 16 patients.

Sujets	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Dépendance	8	3	6	2	4	2	5	9	7	6	3	6	9	5	8	4
Anxiété	4	5	5	4	8	6	4	10	2	3	6	7	5	7	8	6

Classement en rang :

Sujets	Dépendance	Rang D.	Anxiété	Rang A.
1	8		4	
2	3		5	
3	6		5	
4	2		4	
5	4		8	
6	2		6	
7	5		4	
8	9		10	
9	7		2	
10	6		3	
11	3		6	
12	6		7	
13	9		5	
14	5		7	
15	8		8	
16	4		6	

L'ordre choisi est croissant. Ensuite, il vous faut calculer la différence de rang :

d = rang dépendance (x) - rang anxiété (y)

Puis d'élever ces valeurs de d au carré.

Sujets	Rang D.	Rang A.	d = Rd - Ra	d^2
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				

Faire la somme de ces valeurs, soit $\Sigma d^2 =$ …… ;

Multiplier par 6 ce résultat : …… .

Calculer $n (n^2 - 1)$, soit …… = …… .

Obtenir l'équation de $\rho =$ …… = …… .

Conclusion : …….

Chapitre 14 : Variables qualitatives et Chi carré.

Plus adapté aux variables qualitatives, le principe de ce test est le calcul de la distance séparant l'effectif observé de l'effectif théorique. On retrouve le niveau d'inférence, le retour au départ de l'échantillon vers la population. Non plus de moyenne théorique ou espérance, il s'agit ici d'effectif théorique puisque nous sommes faces à des variables qualitatives.

Il s'agit bien d'un test paramétrique, nous interpellons notre hypothèse expérimentée face à l'hypothèse nulle, force du hasard.

Sa formule est : $\chi^2 = \sum_{ij} \frac{(O_{ij} - T_{ij})^2}{T_{ij}}$

- O est l'effectif observé,
- T l'effectif théorique,
- i et j indiquent les différentes classes.

En d'autres termes, nous pouvons dire :

- Le χ^2 global d'un tableau à plusieurs cases est égales à la somme des χ^2 par cases ;
- Le χ^2 d'une case est égale au carré de la différence entre l'effectif observé et l'effectif théorique divisé par l'effectif théorique :

$$\frac{(\text{Eff. Obs.} - \text{Eff. Th.})^2}{\text{Eff. Th.}}$$

Un exemple permet de mieux percevoir le fonctionnement de ce calcul : Enquête sur les conduites alimentaires au petit-déjeuner. Un tableau à double entrée distingue le sexe du type d'aliments.

Type / Sexe	Filles	Garçons	Total
Tartines	25	31	56
Céréales	38	19	57
Total	63	50	113

Ce type de tableau est dit tableau de contingence. Il vise à déterminer la relation entre deux variables. Le raisonnement se mène à partir de l'hypothèse nulle. Dans l'exemple, il n'y a pas de différence entre le groupe des filles et celui des garçons. Il s'agit donc de calculer quel effectif théorique aurait chaque case pour valider cette hypothèse nulle.

Le tableau des effectifs théoriques se construit de sorte que les distributions restent identiques, autrement dit en conservant les valeurs à la marge (en bleu dans le tableau).

Type	Sexe	Filles	Garçons	Total
Tartines	Eff Obs.	25	31	56
	EFF Th.	(56x63)/113 = **31,22**	(56x50)/113 = **24,78**	
Céréales	Eff Obs.	38	19	57
	EFF Th.	**31,78**	**25,22**	
Total		63	50	113

Nous pouvons d'emblée observer l'écart existant entre les deux effectifs et que cet écart varie d'une case à l'autre. Le calcul du Chi carré par carré permet le calcul du χ^2 global.

Type	Sexe	Filles	Garçons
Tartines		χ^2 = …… χ^2 = ……	χ^2 = ……
Céréales		χ^2 = ……	χ^2 = ……
TOTAL : …… = ……			

Une fois calculé, il reste à comparer le résultat du Chi carré à sa table de distribution afin de savoir si les données sont favorables ou non à l'hypothèse nulle. Cette table de distribution du χ^2 indique la probabilité en pourcentage de la part du hasard dans les résultats. Cette probabilité autorise un degré de liberté.

Formule : ddl = (L – 1) x (C – 1)
L : nombres de lignes – C : nombres de colonnes

Pour rejeter H0, le χ^2 doit être supérieur au χ^2 lu dans la table[59] du Chi carré pour un risque donné. Nous le savons, pour attester d'une différence statistiquement significative, *p* doit être égal ou inférieur à 0,05. Il est heureux de constater que les principes comme le seuil de confiance restent identiques. Dans notre exemple :

pour un ddl à …… et pour un p à 0,05, la table du χ^2 nous indique : …… .

Le χ^2 est …… au χ^2 lu, l'hypothèse nulle est …… avec …….

Ce raisonnement peut-il être poursuivi pour un p inférieur ?

……

Conclusion de cet exemple : ……

Par contre, le χ^2 n'indique pas le sens de la relation mais il suffit de regarder le sens des écarts entre les effectifs observés et les effectifs théoriques.

Type	Sexe	Filles	Garçons
Tartines	Eff Obs	25	31
	Δ	25 – 31,2 = - 6,22	+ 6,22
	Eff Th	31,2	24,78
Céréales	Eff Obs	38	19
	Δ	+ 6,22	- 6,22
	Eff Obs	31,78	25,22

[59] En fin de balise.

Conclusion n° 2 de l'exemple **:** les garçons mangent plus fréquemment des tartines que les filles (selon leurs déclarations) tandis que les filles mangent plus fréquemment des céréales au petit déjeuner que les garçons.

Le Chi carré peut également se calculer à partir d'une prévalence théorique. Il s'agit de comparer l'effectif observé par l'enquêteur à d'autres valeurs considérées comme théoriques.
Exemple : dans une unité, on constate que sur 47 patients âgés de plus de 70 ans et alités, 12 ont présenté des escarres au bout d'un mois. La prévalence théorique d'apparition d'escarres pour ce type de population est connue et s'élève à 30 %. Comparons donc les deux mesures.

	Présence d'escarres	Absences d'escarres	Totaux
Effectifs observés	12	35	47
Répartition théorique	30 %	70 %	100 %
Effectifs théoriques	(47x30) / 100 = 14,1	32,9	47

Le calcul du χ^2 : $((12 - 14,1)^2 / 14,1) + ((35 - 32,9)^2 / 32,9) = 0,047$.
Le ddl est $2 - 1 = 1$.
Pour $p = 0,05$, la table indique 3,841 pour un degré de liberté de 1.
Le χ^2 trouvé est inférieur au χ^2 de la table, H0 ne peut pas être rejetée pour un risque de 5 %.
Conclusion de l'exemple : la différence entre les deux distributions n'est pas significative. La prévalence observée correspond à la prévalence théorique du nombre d'escarres dans cette catégorie de patients.
Le Chi carré est une approximation de la différence entre des effectifs. Plus l'effectif est faible, moins le χ^2 est fiable. Sa mesure est inutilisable lorsqu'au moins un des effectifs (total ou partiel de la distribution, théorique ou observée) est inférieur à 5. Dans le cas de petits échantillons, certains tests sont envisageables visant à corriger (de manière relative) cette limite. La correction de Yates propose un calcul du χ^2 en diminuant de moitié l'écart entre la distribution observée et la distribution théorique. Contrainte supplémentaire : cette correction n'est utilisable que dans le cas d'une distribution d'un caractère qualitatif à deux cases.

$$\text{Formule par case : } \chi^2 = \frac{(\text{(Eff. Obs.} - \text{Eff. Th.)}) - 0,5}{\text{Eff. Th.}}$$

Chapitre 15 : Estimation ou intervalles de confiance.

La moyenne reste l'outil principal de la statistique : la moyenne de mon échantillon correspond à la moyenne de ma population si l'hypothèse nulle a été rejetée et que mon échantillon est représentatif de la population. Ainsi, la moyenne d'un échantillon aléatoire permet d'estimer la moyenne vraie (μ) de la population.

L'estimation est le problème inverse de l'échantillonnage ; c'est-à-dire connaissant des renseignements sur un échantillon, on cherche à déduire des informations sur la population totale. Le principe est simple : la moyenne de mon échantillon est la meilleure estimation de la moyenne de la population. Ce type d'estimations est directement lié au choix et à la qualité de l'échantillon.

Ainsi, nous voudrions estimer la précision de cette moyenne, c'est-à-dire donner une marge d'erreur ou intervalle de confiance. Les tables de la loi normale ou de *Student* sont, à nouveau, utilisées pour estimer ces intervalles de confiance. En général, nous adopterons l'intervalle de confiance à 95%, soit à 2σ. Notons que 5 % c'est le niveau de probabilité minimum requis pour accepter l'absence du hasard dans les tests d'hypothèse.

$$\mu = \overline{X} \pm 2\sigma\left(\overline{X}\right)$$

Cette manière détermine donc les bornes entre lesquelles se situent la moyenne inconnue, vraie, avec un risque maximal accepté de 5 %.

Le cas le plus courant est celui où ni la moyenne, ni l'écart-type de la population ne sont connus. Le choix de la loi de référence pour l'estimation de l'intervalle de confiance suit les mêmes règles que précédemment et est donc conditionné par la taille de l'échantillon.

1. L'échantillon est d'un effectif inférieur ou égal à 30 :

   Sur une variable quantitative, l'opération se porte donc sur la moyenne.

   Soit P la population : μ la moyenne est inconnue et σ l'écart-type est inconnu.

   Soit un échantillon e : X_e la moyenne est connue, σ_e l'écart-type est connu et n l'effectif est connu et inférieur ou égal à 30.

   On fera donc référence ici à la loi de *Student* en la suivant à n-1 degrés de liberté (ddl).

   L'intervalle de confiance de la moyenne de la population, avec le coefficient de confiance $2\sigma(t)$-1, lu dans la table de Student à n-1 degrés de liberté est :

$$IC\ (\mu) = t_{n-1} \frac{\sigma_e}{\sqrt{n}}$$

Exemple :

On veut estimer la moyenne vraie des performances au 25 mètres brasse pour des élèves de 6ème année. Pour cela, on étudie un échantillon de 28 élèves en mesurant leurs performances.

On obtient une moyenne X_e = 45 s et un écart type σe = 10 s.

A partir de ces données, on veut estimer l'intervalle de confiance dans lequel la moyenne vraie a 95 % de chances de se trouver. L'échantillon est de 28 (n) et donc inférieur à 30, la formule fait bien référence à la loi de *Student*.

Dans la table, on lit pour p = 0,05 (95 %) et pour un ddl à (n-1) soit **……** une valeur de t = **……** .

$IC_{95\%}$ = …… = ……

$\mu_{IC95\%}$ = […… ; ……]

On peut bien entendu augmenter la précision de cet intervalle de confiance en passant à 3σ, soit 99 %. Il suffit alors de changer de colonne de lecture pour le t de Student. Pour notre exemple, dans la table, on lit pour p = 0,01 (99 %) et pour un ddl à …… une valeur de t = …… .

$IC_{99\%}$ = ……

$\mu_{IC99\%}$ = […… ; ……]

2. L'échantillon est d'un effectif supérieur à 30 :

   Le principe de l'inférence est identique, seule l'équation et surtout la loi de référence changent. Il convient désormais d'utiliser la loi normale. Pour la facilité, nous utiliserons la table de la loi normale réduite simplifiée, comme nous l'avons déjà fait pour les tests paramétriques. Précisons toutefois qu'au plus la table est réduite, centrée et simplifiée, outre le fait que les équations changent, la précision et donc la qualité de l'inférence sont inversement proportionnelles à la simplification opérée. Ici, nous ferons l'expérience avec les deux tables de contrôle.

   Reprenons donc notre exemple en changeant un paramètre : on veut estimer la moyenne vraie des performances au 25 mètres brasse pour des élèves de 6ème année. On étudie un échantillon de 58 élèves en mesurant leurs performances. On obtient X_e = 45 s et σe = 10 s.

   - Alors l'intervalle de confiance de la moyenne de la population, avec le coefficient de confiance 2σ (μ_α ou Z), lu dans la table de la loi normale est :

$$IC\ (\mu) = \mu_\alpha \frac{\sigma_e}{\sqrt{n}}$$

$\sigma_{e\,/}\sqrt{n}$ = …… = …… nous donne le μ_α ou Z à reporter sur la table de la loi normale pour un risque de 5 % , on trouve une valeur à …… .

IC (μ) = …… = …… .

$\mu_{IC95\%}$ = …… $\pm$ …… = […… ; ……]

- Alors l'intervalle de confiance de la moyenne de la population, avec le coefficient de confiance à 95 %, lu dans la table de la loi normale réduite simplifiée est : $\varepsilon = 1{,}96$.

$$IC\ (\mu) = \varepsilon \frac{\sigma_e}{\sqrt{n}}$$

IC (μ) = …… = ……

μ_{IC} = …… $\pm$ …… = […… ; ……]

3. Il est possible d'opérer le même type d'inférence sur les résultats obtenus à la mesure d'une variable qualitative. Il s'agit donc de calculer l'intervalle de confiance d'une proportion plutôt que d'une moyenne. De ce fait même, la technique est plus relative, et est à utiliser avec précaution. Elle utilise la table de l'écart réduit.

   Pour rappel, pour un risque consenti de 5 chances sur 100, la valeur du coefficient ε est de 1,96.

   Condition d'application : $n \geq 5$.

   Soit f (fréquence) = p/100 et g = (1 - p)/100

   $$IC\ (p) = \varepsilon \sqrt{(fg)/n}$$

   Exemple : L'enquête obtient à l'expression d'un item qualitatif (l'utilisation d'un « truc génial ») une proportion de 48 individus sur un effectif de 58. Si mon échantillon est représentatif et non biaisé, quel est l'intervalle de confiance à la population ?

   f = 48 /58 = 0,828 ou p = 82,8 %

   IC (p) = 1,96 x $\sqrt{}$ (0,828 x 0,172) / 58 = 0,0969 soit 9,69 %.

   Conclusion : si on s'intéressait à la population, on aurait seulement 5 chances sur 100 de se tromper en affirmant que le pourcentage d'utilisation de ce « truc génial » se situe entre 73,1 % et 92,5 % (arrondi).

Chapitre 16 : Les tables d'usage statistique

1. Table de nombres aléatoires
2. Tables de probabilité statistique :
   - Table des valeurs significatives de *r*
   - La table de la loi normale centrée
   - La table de la loi normale centrée réduite simplifiée
   - La table du t de Student
   - La table du Chi carré

1. Utilisation de la table de nombres aléatoires

Définition: Une table de nombres aléatoires est une table où chacun des chiffres ou chaque séquence de chiffres a la même chance d'apparaître. Il existe un nombre infini de tables de nombres aléatoires, des logiciels peuvent également fournir cette liste qui permet d'effectuer un échantillonnage aléatoire.

En voici deux exemples :

Table de nombres aléatoires n° 1 :

33 398	99 151	11 851	33 167	82 759	90 258	90 776	54 784
14 987	79 632	53 506	03 555	15 037	47 111	09 578	13 101
80 976	67 577	94 022	31 439	59 609	26 832	84 285	03 116
46 657	70 382	63 743	00 661	96 798	74 197	89 595	56 915
13 879	51 502	47 978	**74 805**	**16 625**	34 670	**04 093**	16 116
00 143	96 272	80 163	95 833	38 538	98 352	**19 041**	33 618
98 960	83 982	16 270	38 963	62 385	50 173	**28 417**	31 616
71 448	66 190	90 481	**23 805**	50 642	26 340	00 205	15 855

Méthode :

1°- On numérote les unités d'observation de façon à ce que chaque unité ait un numéro distinct (on utilisera le nombre de chiffres nécessaires pour écrire la plus grande valeur possible) ;

2°- En se fermant les yeux, on choisit une case au hasard et une façon de se déplacer dans la table ;

3°- Pour chaque case parcourue, si le nombre correspond au numéro d'une unité non encore choisie, on sélectionne cette unité ; sinon, on passe à la case suivante ;

4°- On répète l'étape 3 jusqu'à ce que le nombre d'unités désiré soit atteint.

Exemple

Supposons que, dans la classe, on désire former un échantillon aléatoire de 5 individus parmi les 30 inscrits.

- On utilise les numéros de la liste de classe ; le plus grand nombre étant 30, on utilisera donc les deux premières colonnes de chaque case ;
- On pointe une case au hasard (en rouge ci-dessus) et on décide de parcourir la table de gauche à droite, puis de haut en bas ;
- Le nombre 74 ne correspondant à personne, on passe à la case suivante ;
- En répétant le processus, on sélectionne successivement (cases en gras) les individus portant les numéros: 16, 4, 19, 28 et 23.

Table de nombres aléatoires n° 2 :

38094	7683	26971	60063	6867	65763	52808	69691	90664	47609
11717	56032	55309	35726	9903	10069	89715	11071	51193	17054
86191	63735	89809	46931	63645	91580	77658	35568	45045	77233
41432	61537	89134	55914	10278	14696	55400	78305	97292	8670
97548	34918	73594	77139	36406	95949	86495	76497	91732	66578
60198	98076	30591	39126	6551	76556	18548	71056	49784	92171
10363	74503	72398	81994	50061	76799	82437	78169	99320	13953
23573	13217	19338	84614	56855	88115	33143	50682	40379	87088
73032	6601	85783	52874	66870	66928	89138	7140	5853	33479
32258	77230	78984	83338	85097	84330	15579	88980	20283	48290
62806	35813	36117	4643	93546	53597	12292	77717	72256	39622
87418	11407	65203	16305	59912	33132	95518	89060	96896	65730
61219	93457	57886	51029	51147	87760	77099	14694	91067	3258
75823	55426	78426	756	99670	61624	41398	69477	92962	33737
71410	78	10243	40122	72259	38801	19344	27954	32191	97336
7230	885	97912	17025	72467	7379	57386	35360	4973	31922
12198	42814	84754	38518	68212	87339	1164	25080	44045	21110
13788	67563	56868	99560	58499	60334	95544	40860	78562	43767
30374	41738	9721	59145	91059	33008	70671	39907	32544	12330
60459	16318	25432	71851	19949	3846	43209	10682	81623	11925
43257	78484	49392	88424	66961	65491	91164	38767	25966	39459
36708	17478	2816	92642	71462	95880	17382	12903	18569	64365
44129	90481	43456	32119	77852	35303	18171	68444	32599	93007
69169	56776	14894	39103	7058	12896	38922	17078	87682	95179
15625	57767	21338	68178	68201	69978	72861	93312	93998	84645
85675	74765	78083	88616	42755	86481	63590	10288	81365	54065
26569	74085	51085	30786	23833	78940	96140	34638	24441	1483
23018	24069	18904	33924	78690	69364	12397	54840	89363	91371
71713	7882	87079	78533	66437	95083	5690	84107	15270	28118
16755	4283	92637	86198	72160	91556	38863	87413	89488	98836
57581	18020	3948	2627	12931	77793	87686	99550	22178	47671
80790	79482	95018	53246	37309	22286	81448	64886	58626	32277
40077	8726	45150	25604	43902	4998	20120	16277	20993	896
81677	12846	32662	41691	10580	92231	12863	33832	41223	44151
13635	40804	25480	82993	88041	8029	59133	67032	42295	67162

2. Tables de probabilité statistique :

1 - Valeurs significatives de *r* ou table de Bravais Pearson

N P =	0,10 p =	0,05 p =	0,02 p =	0,01
5	.81	.88	.93	.96
6	.73	.81	.88	.92
7	.67	.75	.83	.87
8	.62	.71	.79	.83
9	.58	.67	.75	.80
10	.55	.63	.72	.76
11	.52	.60	.69	.73
12	.50	.58	.66	.71
13	.48	.55	.63	.68
14	.46	.53	.61	.66
15	.44	.51	.59	.64
16	.43	.50	.57	.62
17	.41	.48	.56	.61
18	.40	.47	.54	.59
19	.39	.46	.53	.58
20	.38	.44	.52	.56
21	.37	.43	.50	.55
22	.36	.42	.49	.54
27	.32	.38	.45	.49
32	.30	.35	.41	.45
37	.27	.32	.38	.42
44	.26	.30	.36	.39
47	.24	.29	.34	.37
52	.23	.27	.32	.35
62	.21	.25	.29	.32
72	.20	.23	.27	.30
82	.18	.22	.26	.28
92	.17	.21	.24	.27
102	.16	.19	.23	.25

2 - Table dite de la loi normale centrée de distribution :

Z	*0,00*	*0,01*	*0,02*	*0,03*	*0,04*	*0,05*	*0,06*	*0,07*	*0,08*	*0,09*
0.00	1.00	.992	.984	.976	.968	.960	.952	.944	.936	.928
0.10	.920	.912	.904	.897	.887	.881	.873	.865	.857	.849
0.20	.841	.834	.826	.818	.810	.803	.795	.787	.779	.772
0.30	.764	.757	.742	.741	.734	.726	.719	.711	.704	.697
0.40	.689	.682	.674	.667	.660	.653	.646	.638	.631	.624
0.50	.617	.610	.610	.596	.589	.582	.575	.569	.562	.555
0.60	.549	.542	.542	.529	.522	.516	.509	.503	.497	.490
0.70	.484	.478	.472	.465	.459	.453	.447	.441	.435	.430
0.80	.424	.418	.412	.407	.401	.395	.390	.384	.379	.373
0.90	.368	. 363	.358	.352	.347	.342	.337	.332	.327	.322
1.00	.317	.313	.308	.303	.298	.294	.289	.285	.280	.276
1.10	.271	.267	.263	.258	.254	.250	.246	.242	.238	.234
1.20	.230	.226	.222	.219	.215	.211	.208	.204	.201	.197
1.30	.194	.190	.187	.184	.180	.177	.174	.171	.168	.165
1.40	.162	.159	.156	.153	.150	.147	.144	.142	.139	.136
1.50	.134	.131	.129	.126	.124	.121	.119	.116	.114	.112
1.60	.110	.107	.105	.103	.101	.100	.097	.095	.093	.091
1.70	.089	.089	.085	.084	.082	.080	.078	.077	.075	.073
1.80	.072	.070	.069	.067	.066	.064	.063	.062	.060	.059
1.90	.057	.056	.055	.054	.052	.051	.050	.049	.048	.047
2.00	.046	.044	.043	.042	.041	.040	.039	.038	.038	.037
2.10	.035	.035	.034	.033	.032	.032	.031	.030	.029	.029
2.20	.028	.027	.026	.026	.025	.024	.024	.023	.023	.022
2.30	.021	.021	.020	.020	.019	.019	.018	.018	.017	.017
2.40	.016	.016	.016	.015	.015	.014	.014	.014	.013	.013
2.50	.012	.012	.012	.011	.011	.011	.010	.010	.010	.010
2.60	.009	.009	.009	.009	.008	.008	.008	.008	.007	.007
2.70	.007	.007	.007	.006	.006	.006	.006	.006	.005	.005
2.80	.005	.005	.005	.005	.004	.004	.004	.004	.004	.004
2.90	.004	.004	.004	.003	.003	.003	.003	.003	.003	003
3.00	.003									

3 - Table dite de la loi normale centrée réduite simplifiée :

Probabilité	Valeur		Probabilité	Valeur
0,00001	4,4145		0,15	1,4395
0,0001	3,8906		0,20	1,2816
0,0005	3,4808		0,25	1,1504
0,001	3,2905		0,30	1,0364
0,005	2,8070		0,40	0,8416
0,01	2,5758		0,50	0,6745
0,025	2,2414		0,60	0,5244
0,05	**1,9600**		0,70	0,3853
0,075	1,7805		0,80	0,2533
0,1	1,6449		0,90	0,1257

Table du t de Student

ddl	p = 0,10	p = 0,05	p = 0,02	p = 0,01	p = 0,001
1	6,314	12,706	31,821	63,657	636, 62
2	2,920	4,303	6,965	9,925	31,598
3	2,353	3,182	4,541	5,841	12,924
4	2,132	2,776	3,747	4,604	8,610
5	2,015	2,571	3,365	4,032	6,869
6	1,943	2,447	3,143	3,707	5,959
7	1,895	2,365	2,998	3,499	5,408
8	1,860	2,306	2,896	3,355	5,041
9	1,833	2,262	2,821	3,250	4,781
10	1,812	2,228	2,764	3,169	4,587
11	1,796	2,201	2,718	3,106	4,437
12	1,782	2,179	2,681	3,055	4,318
13	1,771	2,160	2,650	3,012	4,221
14	1,761	2,145	2,624	2,977	4,140
15	1,753	2,131	2,602	2,947	4,073
16	1,746	2,120	2,583	2,921	4,015
17	1,740	2,110	2,567	2,898	3,965
18	1,734	2,101	2,552	2,878	3,922
19	1,729	2,093	2,539	2,861	3,883
20	1,725	2,086	2,528	2,845	3,850
21	1,721	2,080	2,518	2,831	3,819
22	1,717	2,074	2,508	2,819	3,792
23	1,714	2,069	2,500	2,807	3,767
24	1,711	2,064	2,492	2,797	3,745
25	1,708	2,060	2,485	2,787	3,725
26	1,706	2,056	2,479	2,779	3,707
27	1,703	2,052	2,473	2,771	3,690
28	1,701	2,048	2,467	2,763	3,674
29	1,699	2,045	2,462	2,756	3,659
30	1,697	2,042	2,457	2,750	3,646
∞	1.60	1.96	2.33	2.58	3.29

Table du Chi carré

ddl	p = 0,10	p = 0,05	p = 0,02	p = 0,01	p = 0,001
1	2,706	3,841	5,412	6,635	10,827
2	4,605	5,991	7,824	9,210	13,815
3	6,251	7,815	9,837	11,345	16,266
4	7,779	9,488	11,668	13,277	18,467
5	9,236	11,070	13,388	15,086	20,515
6	10,645	12,592	15,033	16,812	22,457
7	12,017	14,067	16,622	18,475	24,322
8	13,362	15,507	18,168	20,090	26,125
9	14,684	16,919	19,679	21,666	27,877
10	15,987	18,307	21,161	23,209	29,588
11	17,275	19,675	22,618	24,725	31,264
12	18,549	21,026	24,054	26,217	32,909
13	19,812	22,362	25,472	27,688	34,528
14	21,064	23,685	26,873	29,141	36,123
15	22,307	24,996	28,259	30,578	37,697
16	23,542	26,296	29,633	32,000	39,252
17	24,769	27,587	30,995	33,409	40,790
18	25,989	28,869	32,346	34,805	42,312
19	27,204	30,144	33,687	36,191	43,820
20	28,412	31,410	35,020	37,566	45,315
21	29,615	32,671	36,343	38,932	46,797
22	30,813	33,924	37,659	40,289	48,268
23	32,007	35,172	38,968	41,638	49,728
24	33,196	36,415	40,270	42,980	51,179
25	34,382	37,652	41,566	44,314	52,620
26	35,563	38,885	42,856	45,642	54,052
27	36,741	40,113	44,140	46,963	55,476
28	37,916	41,337	45,419	48,278	56,893
29	39,087	42,557	46,693	49,588	58,302
30	40,256	43,773	47,962	50,892	59,703