UNIVERSITÉ DE NEUCHÂTEL

INTRODUCTION A LA STATISTIQUE

Horaires et lieux :
Cours : Lundi de 16h15 à 18h, aula des jeunes rives.

Travaux pratiques : jeudi de 8h30 à 10h, salles D61 et D69, av. du 1er Mars 26.
Les TP ont lieu chaque semaine, mais les étudiant-e-s ne doivent les suivre qu'une semaine sur deux. En plus de l'inscription normale aux TP, vous devez inscrire sur Claroline si vous pensez venir à la première ou à la seconde session de TP, dans le dossier TP STAT1 - Travaux pratiques de statistique I à la rubrique **Groupes**. Faites-le le plus rapidement possible.

Contact :
Eric Crettaz, Cédric Jacot
Université de Neuchâtel, Maison d'analyse des processus sociaux (MAPS)
Faubourg de l'Hôpital 27
2000 Neuchâtel

032/718.14.13
cedric.jacot@unine.ch

Examen : écrit de 2 heures, avec documentation, 4 crédits.
Travaux pratiques : évaluation interne, 1 crédit.

Quelques références bibliographiques, parmi les très nombreux manuels existant :

BLÖSS Thierry, GROSSETTI Michael (1999), *Introduction aux méthodes statistiques en sociologie*, Paris : PUF

CHANQUOY Lucille (2005), *Statistiques appliquées à la psychologie et aux sciences humaines et sociales*, Paris : Hachette

DUMOLARD Pierre, DUBUS Nathalie, CHARLEUX Laure (2005), *Les statistiques en géographie*, Paris : Belin

DODGE Yadolah (2006), *Premiers pas en statistiques*, Paris : Springer

FINK Arlene (1995), *How to Analyze Survey Data*, Thousand Oaks: Sage Publications

UNIVERSITÉ DE NEUCHÂTEL

INTRODUCTION : IMPORTANCE DE LA STATISTIQUE EN SCIENCES HUMAINES

i. LA STATISTIQUE A RÉVOLUTIONNÉ LA RECHERCHE EN SCIENCES HUMAINES…

Selon Kendall, grand statisticien du 20ème siècle dont je reparlerai dans ce cours, l'origine de la statistique moderne, au sens d'une récolte de données afin de réaliser des analyses sociodémographiques ou économiques, remonte au 17ème siècle. Mais il faut attendre le 19ème siècle pour que l'utilisation des lois statistiques se développe, notamment le fait de pouvoir généraliser à l'ensemble de la population des calculs effectués sur un sous-ensemble de celle-ci tiré au hasard, c.à.d un échantillon de la population. La statistique inférentielle (2ème partie de ce cours) se développe de façon très marquée à la fin du 19ème siècle et au début du 20ème siècle, avec notamment le développement des mesures d'association et des tests.

L'essor des outils statistiques a permis le développement des méthodes quantitatives dans les sciences humaines : sciences sociales et politiques (sociologie, psychosociologie, science politique,…), en géographie humaine, en psychologie, en sciences de l'information et de la communication, en sciences économiques, etc.

Plus tard, le développement des méthodes quantitatives sera intimement lié au développement des outils de calcul, depuis les premiers ordinateurs à cartes perforées qui prenaient des salles entières jusqu'aux laptops d'aujourd'hui sur lesquels on peut facilement installer un logiciel de statistique. On peut ainsi analyser en tout lieu des données qu'on a produites soi-même, soit en faisant passer un questionnaire à un échantillon « représentatif » de la population, ou produites dans le cadre d'une démarche expérimentale. En outre, il est de plus en plus simple de télécharger en ligne des données mises à disposition par des organismes publics, p.ex. FORS (Université de Lausanne). Ces outils devenant de plus en plus facile d'accès et d'utilisation, cela a permis à un nombre croissant de chercheuses et de chercheurs d'en bénéficier pour leurs analyses et de répondre à un spectre croissant de questions de recherche concernant des phénomènes sociaux et psychologiques.

L'essor des méthodes quantitatives a également été de pair avec le développement des outils d'observation : développement des interviews par téléphone, puis développement de la saisie des données au fur et à mesure de l'entretien sur un laptop (lors d'interviews en face-à-face) ou sur un PC (lors d'entretiens téléphonique depuis une centrale d'appel), développement des interviews par internet, développements d'outils de mesure des réactions et des comportements en laboratoire, p.ex. mesure des mouvements oculaires, mesures des temps de réaction, utilisation d'images virtuelles pour générer des situations expérimentales, pour n'en citer que quelques uns. Le développement d'outils de mesures et de communications que les « sujets » peuvent transporter avec eux a également contribué à la production de données analysables au moyens d'outils statistiques.

Bien sûr, tous ces développements ont été des outils au service des psychologues, sociologues, anthropologues, géographes, spécialistes de la communication, etc. La base des recherches en sciences sociales reste évidemment le raisonnement et la formulation d'hypothèses qu'on cherchera à infirmer ou à confirmer, notamment au moyen d'outils statistiques.

i.1 Les échantillons représentatifs

Un aspect fondamental de la statistique est, comme indiqué plus haut, qu'on peut tirer des conclusions sur l'ensemble d'une population à partir d'un échantillon. On obtient un résultat certes entaché d'une erreur, mais on peut quantifier cette erreur. On peut également dire si les liens observés au sein de cet échantillon, p.ex. entre un comportement et des caractéristiques sociodémographiques, pourraient tout aussi bien être le fruit d'un pur hasard, ou si, au contraire, ils sont statistiquement significatifs.

Le fait de sélectionner aléatoirement des échantillons représentatifs de la population pour pouvoir ensuite décrire ces données puis tirer des conclusions pour l'ensemble de la population d'une ville, d'une région, d'un pays, a complètement révolutionné notre façon de percevoir le monde social et son fonctionnement. Un exemple permettra de mieux se rendre compte de ce changement crucial.

Avec le développement de la société industrielle, une nouvelle question sociale s'est posée, à savoir la question ouvrière. Celle-ci a été une grande préoccupation pour des personnes d'horizons idéologiques et sociaux très différents, du pape Léon XIII dans son encyclique Rerum Novarum de 1891 à Karl Marx, en passant par des romanciers comme Emile Zola ou encore de nombreux philanthropes se préoccupant des conditions de vie sordides des ouvriers qui s'entassaient en milieu urbain. C'est en Angleterre que se développent les premières véritables recherches sur la pauvreté en milieu ouvrier, notamment la recherche de Charles Booth à Londres à la fin des années 1870, suivie par les travaux de Seebohm Rowntree dans la ville de York une vingtaine d'années plus tard. Ces études étaient très fastidieuses et très longues à réaliser, puisqu'il s'agissait de véritables recensements : on essayait d'obtenir des informations sur l'ensemble des familles vivant dans les quartiers ouvriers de la ville. Dans le cas de Booth, des visiteurs de la commission scolaire londonienne fournirent des informations sur l'ensemble des familles vivant dans l'East End, et la collecte de ces informations a pris une année entière – sans parler de la durée nécessaire pour rencontrer tous les parents d'élèves…

Moins de vingt ans après l'étude de Rowntree, Arthur Bowley, qui a fait des études en mathématiques, réalise une enquête sur la pauvreté. Mais il propose un changement fondamental et préconise de sélectionner aléatoirement un échantillon de ménages plutôt que d'interviewer toutes les familles concernées, notamment parce qu'il avait été contacté par un petit groupe de personnes ayant à la fois des motivations philanthropiques et un budget limité. Sur la base d'une liste des rues classées par ordre alphabétique, il sélectionne au hasard un certain nombre d'adresses. En 1915, cela permet de réaliser une étude comparant la situation dans quatre villes à un bien moindre coût financier et avec bien moins de personnes à interroger. Aujourd'hui, des institutions internationales comme l'OCDE, Eurostat (l'office statistique de l'Union Européenne), le US Census Bureau effectuent très régulièrement de

telles études dans un grand nombre de pays pour suivre l'évolution de la pauvreté monétaire, et de nombreux/ses chercheurs/ses en sciences humaines utilisent ces données pour leurs analyses.

Outre la quantification de phénomènes psychosociaux, l'autre grand but de ce genre d'études est de pouvoir tirer des conclusions en termes de causalité. Toutefois, dans ce genre d'approches, ce que la statistique permet de mesurer, ce sont des corrélations ; le fait d'aboutir à des conclusions concernant les causes d'un phénomène dépend du raisonnement du chercheur. L'un des pères fondateurs de la sociologie parlait de la méthode des « variations concomitantes », c.à.d le fait d'observer, au moyen de données récoltées auprès d'un grand nombre d'individus, si deux phénomènes varient de façon simultanée, afin d'obtenir une expérimentation indirecte d'un lien de cause à effet, lorsqu'on observe des variations qui vont systématiquement dans le sens prédit par nos hypothèses de recherche.

i.2 La méthode expérimentale

Cette approche des phénomènes psychosociaux est également très répandue dans les sciences humaines, et elle utilise systématiquement des outils statistiques pour tirer des conclusions, notamment pour établir qu'il y a une différence significative entre différents groupes de « sujets » participant à une expérience.

Le but de l'approche expérimentale est de montrer que, dans une situation définie, certains comportements ou certaines attitudes sont prévisibles, du moins très probables. L'idée est de manipuler l'un des éléments (ou plusieurs éléments) de la situation expérimentale pour voir si le comportement prédit se produit et de comparer ce qui s'est passé dans une autre ou plusieurs autres situations. Un point absolument fondamental de cette approche est qu'il faut avoir des hypothèses testables, c.à.d pouvoir se baser sur des résultats observables, donc mesurables. Il peut s'agir de mesures objectives car en partie physiologiques, p.ex. des mouvements oculaires, des temps de réaction, le rythme cardiaque, mais également des réalités plus floues et complexes comme des comportements, des opinions exprimées, des images mentales, etc.

Un autre aspect absolument fondamental consiste dans le fait qu'on puisse manipuler une des « variables explicatives » de la situation, une des conditions de traitement. Cela nécessite, entre autre choses, de disposer d'une procédure expérimentale pré-établie afin que l'expérience soit valide. Il ne faut pas improviser, car cela ajoute des « variables » au problème, p.ex. la réaction du chercheur/de la chercheuse à une question posée par un des participants.

Au final, on se retrouve avec les informations suivantes : plusieurs groupes de sujets, similaires en tous points, ou du moins les plus similaires possibles, se sont retrouvés dans des situations ou un ou plusieurs paramètres que les chercheurs/ses maîtrisent ont été modifiés, et on a observé des réactions différentes qu'on a mesuré au moyen d'électrodes, de questionnaires, d'une interview de debriefing après l'expérience, au moyen d'images vidéo prises pendant l'expérience, au moyen d'un « eyetracker », etc. Il s'agit de savoir si ces différences sont suffisamment marquées pour qu'on puisse en tirer des conclusions en termes de causalité.

L'approche expérimentale s'est également développée à la fin du 19ème siècle. En psychologie, l'Allemand Wilhelm Wundt fut le premier à fonder un laboratoire de psychologie expérimentale à la fin des années 1870. L'approche expérimentale se développa de façon beaucoup plus marquée au 20ème. Dans le domaine de la psychologie sociale, quelques expériences ont profondément marqué les sciences humaines au sens le plus large. Une des plus connues est sans aucun doute l'expérience réalisée à Yale au début des années 1960 par le psychologue American Stanley Milgram pour mieux comprendre la soumission à l'autorité.

Des individus recrutés au hasard par le biais de petites annonces se rendaient au laboratoire de Milgram, et on leur faisait croire qu'il s'agissait d'une étude portant sur les mécanismes d'apprentissage. On procédait à un tirage au sort truqué et les participants se voyaient attribuer le rôle d'enseignant, alors que les autres participants (en fait des comparses de Milgram qui se faisaient passer pour des participants) se voyaient attribuer le rôle d'élève. L'enseignant avait une liste de mots à associer devant lui, et il devait la faire apprendre à l'élève, et l'expérimentateur lui explicait qu'il/elle devrait administrer un choc électrique à chaque mauvaise réponse, pour voir si les sanctions sévères améliorent l'apprentissage. Bien sûr, tout cela était factice, l'élève se trouvant dans une autre pièce ne recevant pas de chocs électriques, mais il/elle poussait des cris de souffrances que l'enseignant entendait. Les résultats de Milgram dépassèrent l'imagination : les 2/3 des participants allaient jusqu'au bout des chocs électriques (de plus en plus violents et ils/elles le savaient) malgré les cris de douleurs et le fait que l'élève demandait pitié. Lors d'un debriefing, Milgram constata que la plupart des gens étaient allés jusqu'au bout car ils/elles ne se sentaient pas vraiment responsables de leurs actes (ce que Milgram appela l'état « agentique »). Plusieurs expériences furent réalisées, en faisant varier certaines conditions, p.ex. la proximité avec l'élève, et le degré de soumission variait fortement d'une condition à l'autre.

Le but de l'approche expérimentale est d'aborder la causalité de façon directe, car tout est (plus ou moins) identique dans les diverses situations expérimentales, sauf le paramètre que les chercheurs/ses font varier. Contrairement à l'approche présentée plus haut, on travaille avec un petit nombre d'individus, et comme indiqué, on cherche à établir de façon directe le lien entre une cause (la condition expériementale manipulée par les chercheurs/ses) et l'effet (une réaction physiologique, comportementale, etc.), alors que cela n'est pas possible avec une enquête réalisée auprès d'un grand échantillon de la population.

La statistique joue un rôle fondamental dans l'approche expérimentale. En effet, l'une des préoccupations majeures est de voir si les différences observées et mesurées entre plusieurs groupes de sujets sont plutôt le fruit du hasard, ou si au contraire elles sont statistiquement significatives, et cela est particulièrement important car ces groupes contiennent peu d'individus. On va p.ex. comparer les temps de réaction moyens, l'augmentation du rythme cardiaque chez les sujets, le pourcentage des sujets de chaque groupe qui donnent une certaine réponse ou qui adoptent un certain comportement, etc. Pour cela, on utilise des tests de significativité divers et variés, après avoir correctement décrits les résultats observés. Là aussi, les progrès réalisés au 20ème siècle en statistique ont profondément modifiés l'analyse des comportements, des attitudes, des affects, etc.

ii. CES PROGRES ONT EGALEMENT EU UN IMPACT MAJEUR SUR BIEN DES ASPECTS DE LA VIE QUOTIDIENNE

La politique a été profondément affectée par le développement des sondages après la Deuxième Guerre Mondiale. De nos jours, il n'y a pas de campagne électorale sans sondages d'opinions, ni de journée électorale sans sondages de sortie des urnes.

Très peu de produits que vous achetez dans un supermarché sont mis en rayons avant d'avoir été goûtés ou testés par un échantillon de consommateurs, et des tests de type expérimental sont également réalisés dans ces supermarchés pour voir s'il y a des différences de traitement des clients.

Très peu de publicités sont affichées dans la rue ou diffusées à la télévision sans avoir été préalablement montrées à un échantillon d'individus, et les spécialistes de la communication recourent également aux méthodes expérimentales.

Or toutes ces approches reposent sur la statistique : maîtriser celle-ci, c'est donc disposer des outils permettant d'analyser ces méthodes devenues si familières et pourtant si mal comprises par la population, comme en témoignent les débats assez creux qui surgissent à chaque fois qu'un sondage d'opinion ne prédit pas correctement le résultat d'une votation/d'une élection.

Dans ce qui suit, je vais beaucoup parler de l'approche reposant sur un échantillon « représentatif » de la population ; toutefois, la méthode expérimentale ne sera pas délaissée.

UNIVERSITÉ DE NEUCHÂTEL

COLLECTER ET PREPARER LES DONNEES

1. NOTIONS DE BASE

Un des apports fondamentaux de la statistique c'est qu'on peut étudier une population en se focalisant sur un **échantillon**, c'est-à-dire sur un sous-ensemble de cette population.

Les variables (c.à.d. les caractéristiques « mesurables » de la population étudiée) sont indiquées par des lettres majuscules **:** souvent dans ce cours nous utiliserons les lettres X, Y. Les différentes valeurs que peut prendre une variable sont les modalités. Pour désigner la valeur de la réalisation de X pour le $i^{ème}$ individu de l'échantillon, on parle aussi d'observation, on utilisera la notation x_i. Une variable **indépendante** est en fait une variable explicative et une variable **dépendante** est une variable expliquée.

Généralement on utilise la notation **n** pour désigner la taille de l'**échantillon** et N pour la taille de la population.

On distingue généralement deux grands types de variables en sciences humaines:

A. Les variables **qualitatives** qu'on peut classer en deux sous-groupes :

- Les variables nominales ou catégorielles (pas de distance ni de hiérarchie, mais simplement l'appartenance à une catégorie).
- Les variables ordinales (il existe un ordre hiérarchique entre les catégories).

Parmi les variables qualitatives, le nombre de réponses possibles, c.à.d de modalités, peut fortement varier. On peut avoir des variables dichotomiques. A l'opposé, on peut avoir un très grand nombre modalités, p.ex. la profession exercée par le répondant. Dans ces cas-là, on recourt souvent à des nomenclatures.

B. Les variables **quantitatives,** elles aussi divisées en deux sous-groupes :

- Les variables intervalles (échelles d'intervalles)
- Les variables quotient (rapports), ou le 0 a une signification et signale l'absence de quelque chose (p.ex. 0 francs = pas d'argent).

Une distinction importante doit encore être signalée : il existe des **variables continues** qui peuvent prendre n'importe quelle valeur (p.ex. le salaire) et des **variables discrètes**, c.à.d dont l'étendue des valeurs possibles est dénombrable (p.ex. le nombre d'enfants).

Pour choisir une méthode d'analyse appropriée, il faut commencer par définir une question de recherche, ensuite on détermine quelles sont les variables dépendantes et indépendantes et si ces variables sont nominales, ordinales ou quantitatives.

UNIVERSITÉ DE NEUCHÂTEL

2. ECHANTILLONAGE, PONDERATION ET REPRESENTATIVITE

2.1. Population et échantillon

Définir la population de référence constitue une première étape très importante dans l'optique de définir le tirage aléatoire d'un échantillon et les conclusions de l'étude ne pourront porter que sur cette population de référence et les unités statistiques qui la composent.

Un **dénombrement complet** (relevé exhaustif) de la population est très compliqué, très onéreux et très long, en général. Le recensement fédéral de la avait lieu tous les 10 ans en Suisse. Ceci est évidemment un exemple extrême ; si la population de référence n'est pas trop grande ou très hétérogène, cela peut valoir la peine de réaliser un relevé exhaustif. L'avantage d'un dénombrement complet est bien entendu que les paramètres de la population sont connus, alors que pour un échantillon il faut faire des estimations qui sont entachées d'une certaine erreur.

Un **échantillonnage aléatoire** est basé sur des règles de sélection précises pour tirer au sort un sous-groupe de la population. Pour les études à caractère scientifique on définit des règles fixes de sélection ; il existe plusieurs types d'échantillons, p.ex. un échantillon aléatoire simple (on tire au hasard un certain nombre d'éléments de la liste) ou des tirages plus complexes, comme l'échantillonnage stratifié (on définit des strates, p.ex. régionales, et ensuite on effectue un tirage pour chaque région), ou encore des tirages à plusieurs degrés (p.ex. on sélectionne des ménages, puis un adulte par ménage).

La base de la plupart des processus aléatoires d'échantillonnage repose sur le fait d'avoir à disposition une **liste exhaustive** des éléments de la population de référence. En outre, l'autre élément absolument central réside dans le fait d'avoir **la même probabilité d'être sélectionné** ou non. Si cette probabilité varie d'une unité à l'autre, on doit généralement effectuer une *pondération.* En outre, il se peut que le processus de sélection et la réalisation de l'étude entraînent des biais, si les non-réponses, les refus, etc. ne sont pas répartis de façon homogène dans la population (p.ex. les personnes de moins de 30 ans ont moins participé que la population de plus de 30 ans).

La notion de pondération est simple à comprendre. Si l'on constate qu'il y a eu un biais au moment de la sélection aléatoire, car certains groupes de la population sont sous-représentés (sachant le poids que ces individus ont dans la population, ils devraient être plus nombreux dans l'échantillon), on va donner plus de poids à ces répondants. A contrario, on va donner moins de poids aux répondants appartenant à des groupes sociaux sur-représentés. Si le principe est relativement simple, le calcul des facteurs de pondération est plus complexe, et la plupart des bases de données fournies par les organismes de recherche contiennent des facteurs de pondération « prêts à l'emploi ».

2.2. Les échantillons « représentatifs »

Les échantillons aléatoires constituent la seule garantie que l'on puisse tirer des conclusions concernant la population de référence sur la base des caractéristiques de l'échantillon, du moins dans les limites de certaines erreurs statistiques.

En principe c'est la seule dimension vraiment scientifique du terme « représentativité ». En pratique, par contre, on utilise ce terme à tort et à travers. Il ne s'agit pas d'un critère de qualité bien défini.

Pour juger de la qualité d'une procédure d'échantillonnage, il faut au moins disposer d'informations exactes sur : la population de référence, le processus de sélection, les valeurs manquantes, les non-réponses, la distribution de certaines valeurs structurelles et les instruments utilisés.

Si l'on a une approche scientifique, on ne devrait pas se contenter de constater que certaines caractéristiques de l'échantillon sont réparties de la même manière que dans la population de référence et d'en déduire que c'est le cas pour toutes les caractéristiques. Il faut donc se méfier de la méthode des quotas, mais aussi des échantillons aléatoires contenant des non-réponses qui peuvent générer une distorsion.

UNIVERSITÉ DE NEUCHÂTEL

DECRIRE ET COMPRENDRE LA REPARTITION DES DONNEES

3. STATISTIQUE DESCRIPTIVE ET REPRÉSENTATIONS GRAPHIQUES

3.1. Mesures de tendance centrale : moyenne, mode, médiane

On indique un point autour duquel les observations se concentrent. Les mesures les plus couramment utilisées sont la moyenne arithmétique, le mode et la médiane.

La **moyenne arithmétique** est la somme des observations divisée par le nombre d'observations. En langage mathématique on note :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Le **mode** est la valeur d'une variable qui a la plus grande fréquence. Notons qu'il existe des cas où l'on a une distribution bimodale (2 valeurs sortent le plus fréquemment), voire multimodales.

La médiane est la valeur qui divise *la série ordonnée des observations* en deux groupes de même taille ; il s'agit de l' « observation du milieu ». Autrement dit, la moitié des observations sont en dessous de cette valeur, et l'autre moitié en dessus. Si n, la taille de l'échantillon, est paire, alors la médiane est la moyenne arithmétique des deux valeurs centrales. Si n est impaire, alors c'est plus simple puisqu'on a une valeur centrale.

De manière générale, la moyenne arithmétique est très influencée par les valeurs extrêmes, ce qui n'est pas le cas de la médiane. Si la répartition des points est symétrique, alors la médiane et la moyenne se confondent.

3.2. Quantiles, variance et dispersion des données

En fait on peut généraliser l'idée de la médiane : il existe des points qui divisent l'échantillon en plusieurs groupes de « taille » égale : ce sont les quantiles.

On parle de quantile d'ordre α, avec α compris entre 0 et 1 (entre 0 et 100%). Il y a une proportion α des observations en dessous de ce point et une proportion 1-α en dessus.
Les quantiles les plus utilisés en sciences sociales sont les quartiles, les quintiles et les déciles.

Prenons les **quartiles** (définir des groupes contenant ¼ des observations) :

- q_1 qui est le quantile d'ordre ¼, c'est-à-dire la valeur telle que ¼ des observations sont en dessous et ¾ en dessus.
- q_2, c'est-à-dire que 2/4 des observations sont en dessus et 2/4 en dessous: il s'agit en fait de la médiane.
- q_3, ¾ des observations sont en dessous, ¼ en dessus.

Les **quintiles** sont les points qui divisent l'échantillon en groupes de 20% des observations (soit en 5 groupes), et les **déciles** en 10 groupes de 10%.

La **variance** mesure la dispersion des observations Il s'agit *de la moyenne arithmétique des écarts à la moyenne* (élevés au carrés).

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

On peut montrer que cela équivaut à :

$$s^2 = (\frac{1}{n}\sum_{i=1}^{n}x_i^2) - \bar{x}^2$$

L'**écart-type** est la racine carrée de la variance empirique, et on le note s.

D'autres indications de la variation: le maximum, le minimum et l'étendue empirique, soit la différence entre le maximum et le minimum. Cette mesure est très sensible aux valeurs extrêmes, en toute logique.

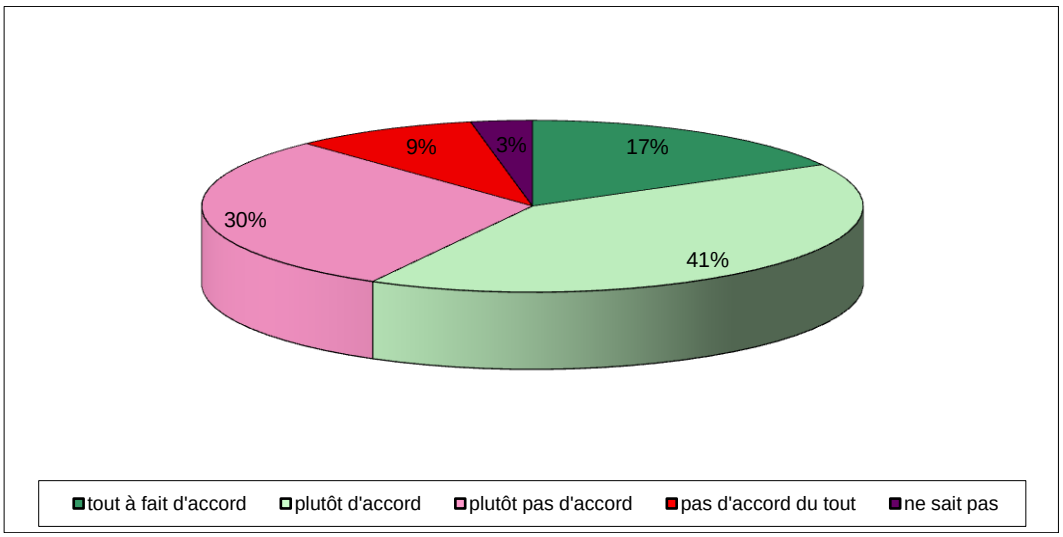
Autre élément, l'**étendue interquartile** (qu'on abrège souvent IQR), qui est en fait tout simplement l'écart entre le troisième et le premier quartile, IQR= q_3 - q_1. C'est une indication intéressante que l'on utilise pour faire un graphique assez utile qu'on appelle un **box-plot** (nous y reviendrons).

3.3. Les représentations graphiques

Le **diagramme en bâton** est utilisé pour les variables qualitatives ou quantitatives discrètes. Chaque bâton est proportionnel au pourcentage / à la fréquence d'une réponse. C'est un type de diagramme qui est très couramment utilisé. On peut aussi y recourir pour les questions avec plusieurs réponses possibles ou pour toute autre situation où le total ne fait pas 100%.

L'**histogramme** repose sur le même principe, il représente la distribution d'une variable quantitative continue. Il faut ordonner les observations et ensuite déterminer un certain nombre de **classes.** On devrait avoir des **classes d'une taille égale** si l'on utilise des barres de même largeur ou alors il faut avoir des barres de „taille" variable (pas pratique à faire sur PC). En outre, n'oubliez pas que souvent la barre la plus à droite du graphique est souvent une classe ouverte (p.ex. 100'000 francs et plus dans un graphique sur la distribution des revenus).

Enfin, un graphique relativement peu utilisé dans la recherche, mais courant dans les sondages d'opinion p.ex. est le **pie-chart ou diagramme circulaire.** C'est un graphique très simple : chaque tranche du gâteau a une taille proportionnelle au pourcentage qu'il représente. Ce graphique ne convient pas pour les questions multiréponses !



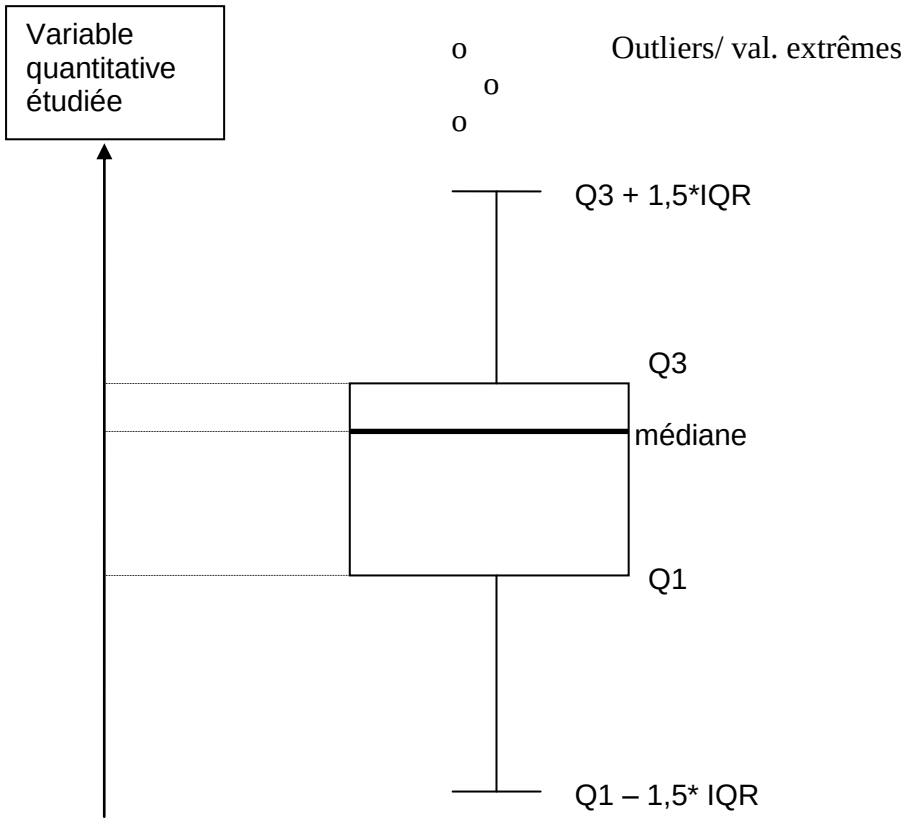
Dans les sondages, on souhaite généralement qu'une **majorité se dégage de l'enquête**, sur le modèle du vote. L'essentiel est d'avoir un chiffre qui puisse figurer dans le titre de l'article, s'il le faut on peut regrouper des catégories de réponses.

Le **box-plot** / boîte à moustaches est un graphique qui a comme qualité de bien résumer l'information. En un coup d'œil on a des informations sur la tendance centrale, sur la dispersion des données, voire sur la symétrie des données et on peut facilement comparer des groupes d'observations.

Le principe est le suivant :

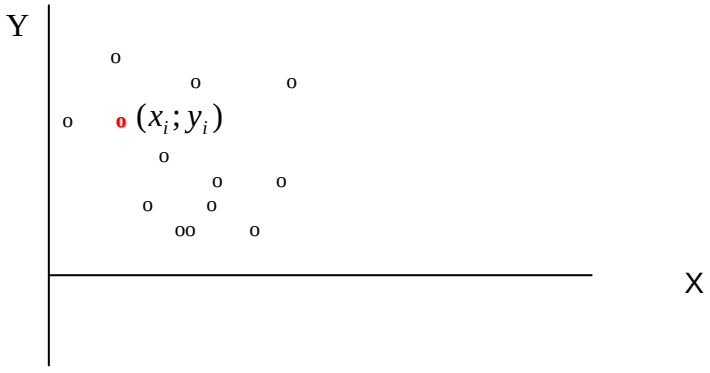
On dessine une **« boîte à moustaches ».** La boîte est définie par le quartile supérieur q3 et le quartile inférieur q1. Dans la boîte on a donc la moitié des observations. A l'intérieur de la boîte on indique la médiane. Ensuite on dessine les moustaches dont la longueur correspond à 1,5 fois l'étendue interquartile.

Voici un exemple de box-plot réalisé selon les principes énumérés (on trouve parfois aussi des graphiques similaires à des box-plots, mais avec les « moustaches définies différemment, p.ex. le décile supérieur et le décile inférieur) :



Les points se situant au-delà des extrémités des « moustaches » sont qualifiés de **valeurs extrêmes** (en anglais : outliers). Les outliers sont des valeurs se trouvant soit en dessus de $q_3 + 1{,}5IQR$ ou en dessous de $q_3 - 1{,}5IQR$; parfois on s'intéresse aux valeurs vraiment extrêmes et on peut les définir au moyen de 3 fois l'écart interquartile au lieu d'une fois et demie. Certains utilisent parfois la moyenne et 3 fois l'écart-type, c.à.d $\bar{x} \pm 3s$.

Le diagramme de dispersion est très utilisé pour comparer deux variables quantitatives. Le principe est très simple: on représente des couples de points $(x_i ; y_i)$.



3.5. Pourcentages, proportions et description des évolutions

Jongler avec des pourcentages paraît être quelque chose d'évident. Pourtant beaucoup d'erreurs sont commises.

On sait que les pourcentages ne sont pas toujours bien compris dans l'ampleur qu'ils représentent. Souvent il est plus parlant, surtout dans des publications non scientifiques, de dire qu'environ un répondant sur sept est d'accord, plutôt que 14,3%. En outre les pourcentages faibles sont souvent peu intuitifs. Donc, il est souvent plus parlant de décrire les résultats avec des proportions arrondies en termes de tiers, de quarts, de 1 répondant sur x, etc.

Comment décrire l'évolution de variables quantitatives (taux de chômage, nombre d'habitants, taille moyenne des ménages, surfaces agricoles dans un pays, etc.) ?

On doit d'abord décrire une tendance générale avant même d'entrer dans les détails : hausse, baisse, stagnation, volatilité ou stabilité, etc.

Il y a trois principaux types de description des évolutions :

i) Pour les pourcentages : l'évolution en **points** de pourcentage, c'est la différence arithmétique entre les %.

ii) l'écart arithmétique : c'est la variation exprimée en % (et non pas en points de pourcentage), sur le modèle : $\frac{\text{situation finale - situation initiale}}{\text{situation initiale}}$

iii) la progression géométrique : qui consiste à diviser la situation finale par la situation initiale, pour voir de « combien de fois » une valeur a augmenté (ou diminué).

Prenons un exemple : le taux de chômage dans un pays passe de 5,3% à 6,3%.

Le taux a augmenté de 1 point de pourcentage (6.3 – 5.3), il a augmenté de 18,9% ($\frac{6.3-5.3}{5.3}$ = 0,188679 ≅ 18,9%) et a donc été multiplié par 1,189 ($\frac{6.3}{5.3}$ = 1,188679).

Notez au passage que la progression géométrique et la progression arithmétique sont liées.

UNIVERSITÉ DE NEUCHÂTEL

4. PROBABILITES

Le raisonnement en sciences sociales (au sens large) est souvent probabiliste : sachant qu'un individu a certaines caractéristiques, il est plus ou moins probable qu'il ou elle soit au chômage, lise le Temps plutôt que l'Illustré, choisisse de faire des études scientifiques plutôt que littéraires, etc.

La probabilité qu'un événement A se produise est noté P(A). La probabilité de n'importe quel événement se calcule de la même façon :

$$P(A) = \frac{\text{nombre de cas ayant la caractéristique étudiée}}{\text{nombre de cas possibles}}$$

4.1. Propriétés des probabilités et notations

P est une probabilité si elle satisfait les propriétés suivantes :

1) $0 \leq P(A) \leq 1$, donc une probabilité est comprise entre 0 et 100%

2) $A \cup B$ signifie que soit l'événement A se produit, soit l'événement B se produit, soit les deux simultanément.

3) $A \cap B$ signifie que les deux événements se produisent simultanément (ou qu'un individu possède les deux caractéristiques)

4) Il existe un événement complémentaire qu'on note $\overline{A}$, qui signifie en fait que A ne se réalise pas. Exemple : ne pas être pauvre. La probabilité de l'événement complémentaire est vaut donc $P(\overline{A}) = 1 - P(A)$.

4.2. Indépendance et probabilités conditionnelles

Soit deux événements A et B. On se demande si la probabilité que B se réalise influence la probabilité que A se réalise.

On note P(A / B) la probabilité que A se réalise sachant que B est réalisé, p.ex. le fait d'être au chômage sachant que le répondant n'a pas de formation post-obligatoire. Il s'agit d'une **probabilité conditionnelle.** Si l'événement B est possible (c'est-à-dire si P(B)>0), on sait que

$$\mathbf{P(A / B)} = \frac{P(A \cap B)}{P(B)}$$

Souvent se pose la question de savoir si deux variables sont indépendantes l'une de l'autre.

C'est un point crucial, et on définit l'indépendance de la façon suivante :

Deux événements sont **indépendants si P(A / B) = P(A),** en d'autres termes si la variable B n'a pas d'influence sur la probabilité que A se réalise.

On peut montrer que si A et B sont des événements indépendants,
$P(A \cap B) = P(A) * P(B)$,

en effet, compte tenu de la formule de calcul d'une probabilité conditionnelle définie ci-dessus :

$P(A \cap B) = P(A / B) * P(B)$ or si A et B sont indépendant $P(A / B) = P(A)$, donc
$P(A \cap B) = P(A) * P(B)$.

Nous y reviendrons quand nous étudierons un test permettant de savoir si deux variables qualitatives sont indépendantes l'une de l'autre.

5. LES VARIABLES ALEATOIRES : PROBABILITES, REPARTITION ET DENSITE

Une variable aléatoire est une variable dont la valeur n'est pas connue avant qu'elle soit observée. Nous en distinguerons deux types : les variables (quantitatives) aléatoires discrètes et continues.

5.1. Les variables aléatoires discrètes

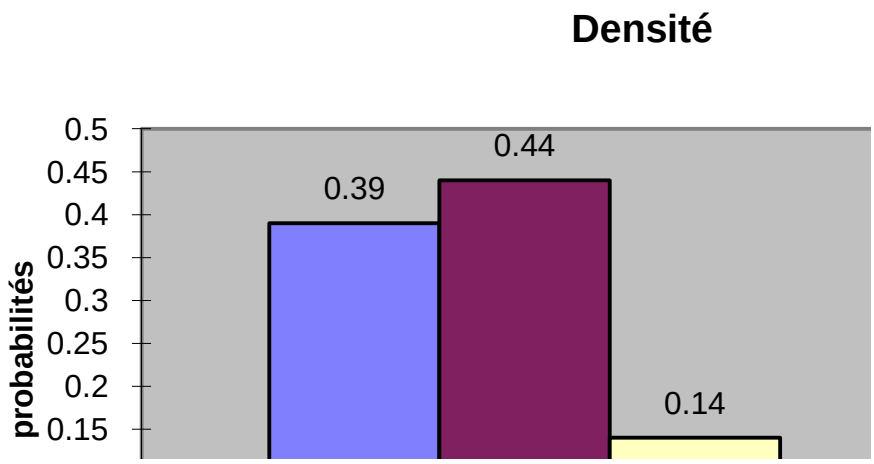
Il s'agit de variables dont chaque valeur a une probabilité strictement positive ou nulle. On écrit pour la variable aléatoire discrète X ayant ses valeurs dans un ensemble fini
$\{\chi_1, \chi_2, \cdots, \chi_k\}$:

$0 \leq P(X = \chi_i) \leq 1$

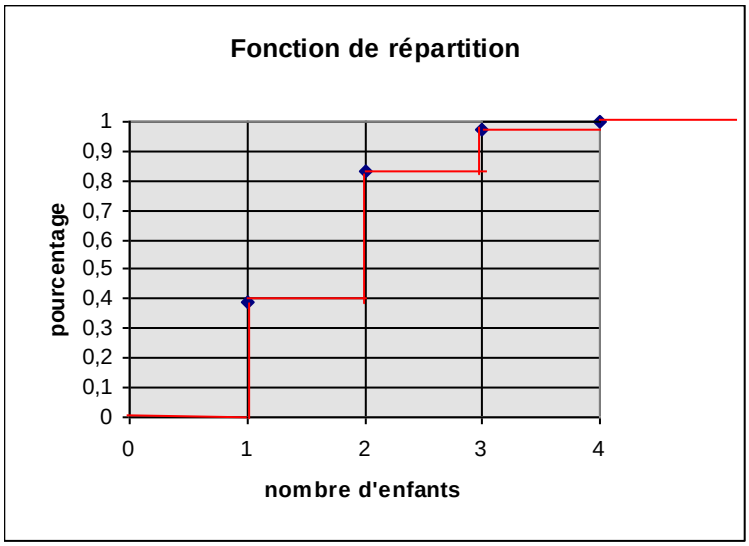
La **loi de probabilité** est une fonction qui associe à chaque valeur χ_i de la variable X la probabilité $P(X= \chi_i)$, par exemple la probabilité d'avoir 2 enfants.

La fonction $F_X(X) = P(X \leq \chi_i)$ est appelée la **fonction de répartition,** et elle varie entre 0 et 1. C'est la probabilité d'avoir une valeur inférieure à un certain seuil, p.ex. la probabilité d'avoir moins de 3 enfants.

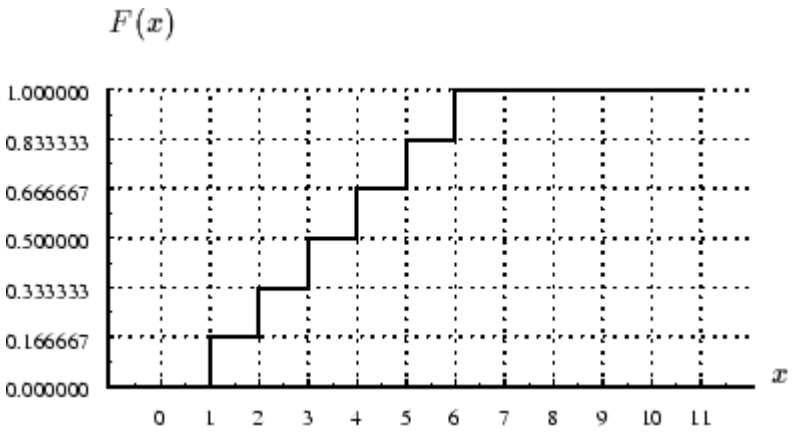
On représente la **densité** de cette variable au moyen d'un histogramme. P.ex. pour le nombre d'enfants :



Et la fonction de répartition est représentée de la manière suivante :



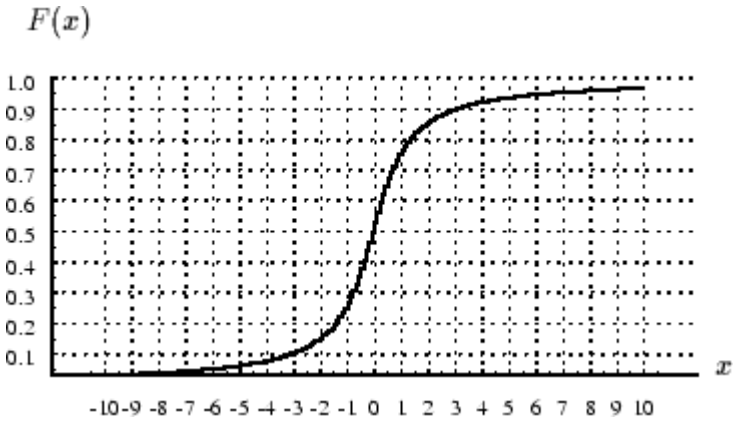
Autre exemple, la fonction de répartition du lancer de dé :



5.2. Les variables aléatoires continues

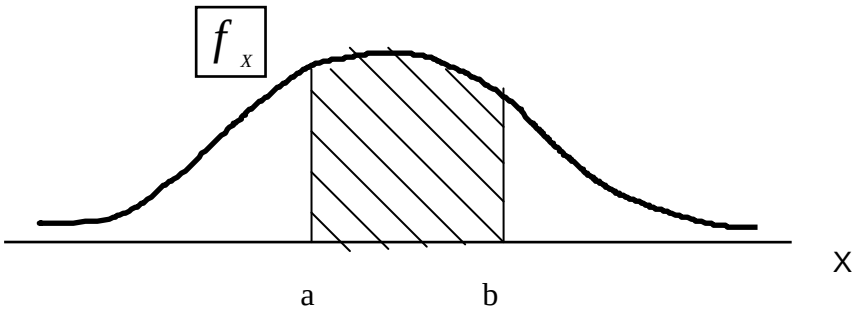
Il s'agit de variables dont chaque valeur prise individuellement a une probabilité **nulle** (ou presque) car il y a très peu de chances qu'un répondant indique précisément cette valeur. Il faut définir des probabilités sur un **intervalle**.

Comme il s'agit de variables continues, la **fonction de répartition** prend la forme suivante :



Cette fonction remplit le même but que l' « escalier » pour les v.a. discrètes, il s'agit en fait d'un escalier avec une infinité de marches. Donc cette fonction de répartition indique, pour chaque valeur x de la variable X, la probabilité d'avoir des valeurs **inférieures à x,** c.à.d $P(X \leq x)$, p.ex. la probabilité de gagner moins de 5500 francs par mois si X est la variable continue « salaire mensuel ».

Pour décrire la variable, on utilise la **fonction de densité** :



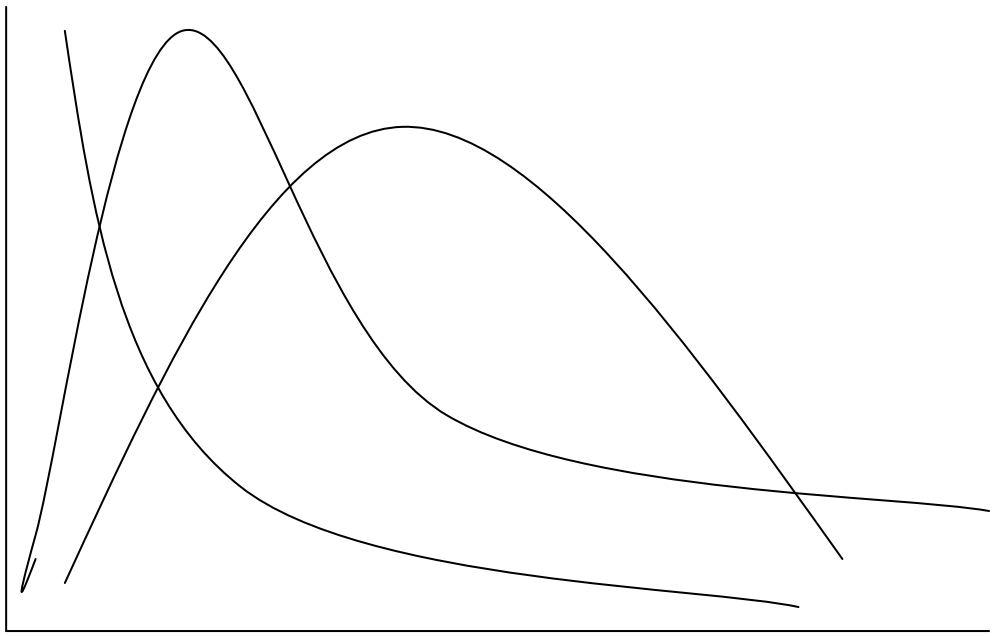
Là où $f_X(X)$ est « élevée », les valeurs de X sont plus fréquentes que là où elle est « basse ».

La probabilité qu'une valeur x se situe entre a et b correspond à l'aire qui se trouve délimitée par a et b et se trouvant sous la courbe (la zone hachurée dans la figure ci-dessus).

5.3. La loi normale

5.3.1 Distribution

La distribution d'une v.a. continue peut prendre toutes sortes de formes :



La densité d'une v.a. suivant une loi normale a la forme d'une cloche symétrique, c'est la fameuse courbe de Gauss.

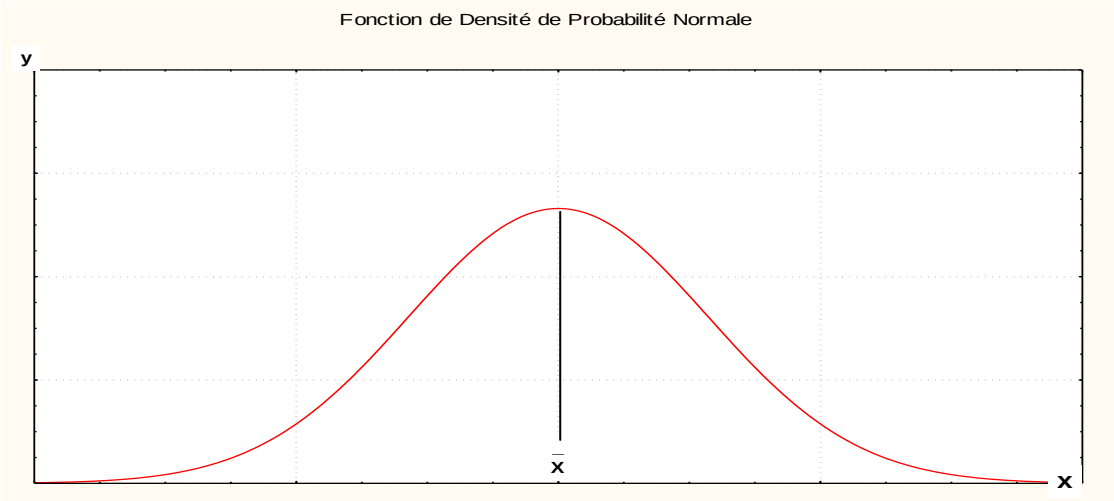
Toute loi normale se caractérise par 2 paramètres : μ et σ^2, une autre notation pour la moyenne et la variance, et on note **$X \sim N(\mu, \sigma^2)$.** On estime ces paramètres de la façon suivante :

- l'estimateur de μ : $\frac{\sum_{i=1}^{n} x_i}{n}$, c'est la moyenne arithmétique,
- l'estimateur de σ^2 : $\frac{\sum_{i=1}^{n} (\bar{x} - x_i)^2}{n}$, c'est la variance empirique.

La formule de la loi normale est la suivante :

$$f_x(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right]$$

Cette fonction a l'allure suivante :



Notons quelques propriétés importantes de cette courbe de Gauss :

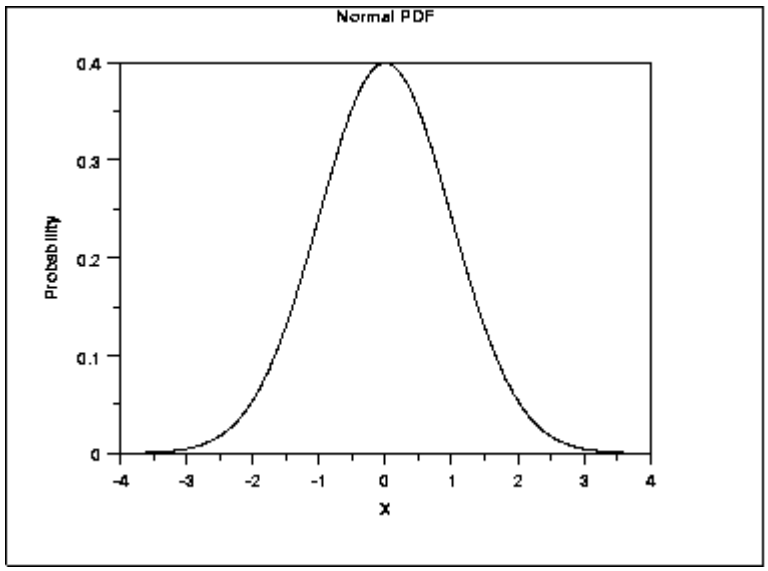
- le maximum de la fonction est la moyenne $\bar{x}$
- les points d'inflexion de la courbe se situent à $\bar{x} \pm s$ ou $\mu \pm \sigma$
- en ce qui concerne l'écart-type, on sait que :

$P(\bar{x} - s \le x \le \bar{x} + s) \approx 68\%$
$P(\bar{x} - 2s \le x \le \bar{x} + 2s) \approx 95\%$
$P(\bar{x} - 3s \le x \le \bar{x} + 3s) \approx 99\%$

5.3.2 Loi normale centrée réduite

Il s'agit de la loi normale ayant pour particularité d'avoir une moyenne nulle et une variance valant 1 : X ~ N (0,1). La formule de la fonction est la suivante :

$f_x(X) = \frac{1}{\sqrt{2\pi}} \cdot \exp(-\frac{x^2}{2}) = \varphi(x)$, qui a l'aspect suivant :



Pour pouvoir utiliser la loi normale centrée réduite pour une variable X ~ N (μ, σ^2), il faut la standardiser, c'est-à-dire transformer cette variable X en une variable Z qui, elle, suit une loi normale centrée réduite :

$$\mathbf{Z} = \frac{X - \mu}{\sigma} \sim \mathbf{N\ (0,1)}$$

Cette transformation z permet d'obtenir ce qu'on appelle parfois le z-score ou score standardisé. On trouve dans des tables la probabilité d'avoir une valeur de Z < qu'une certaine valeur, et qu'on note $\Phi(z)$

Il faut en outre ajouter que :

$\Phi(-z) = \mathbf{1 - \Phi(z)}$, par ce que cette courbe est symétrique autour de 0 (car c'est la moyenne).

Avec ces deux propriétés fondamentales, on peut trouver ces probabilités au moyen des tables usuelles. Dit autrement, n'importe quelle variable suivant une loi normale peut être transformée en une autre variable qui, elle, suit une loi normale centrée réduite. Donc il suffit d'avoir une seule table de probabilités pour toutes les variables suivant une loi normale.

5.3.3 Le théorème central limite TCL

C'est l'un des théorèmes fondamentaux de la statistique. Il dit que si on a un grand nombre de variables aléatoires $x_1, x_2, \ldots, x_n$ indépendantes et identiquement distribuées, alors leur somme $x_1 + x_2 + \ldots + x_n$ suit approximativement une loi normale. On estime que ce théorème est applicable à partir de n = 30.

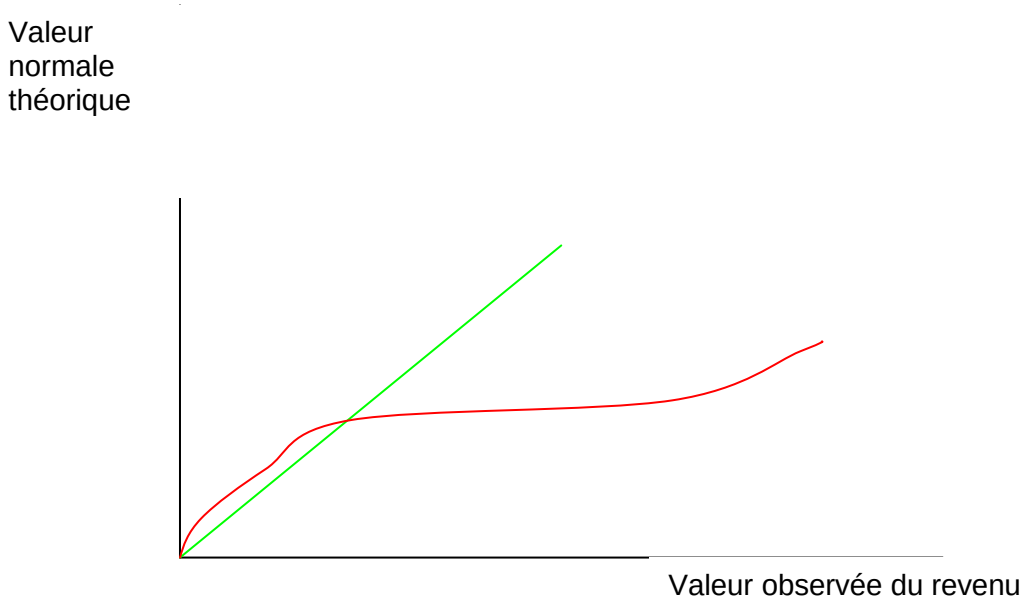
Cela signifie notamment que la moyenne d'un échantillon aléatoire suit une loi normale, c'est-à-dire $\bar{x} \sim N(\mu_{\bar{x}}, \sigma_{\bar{x}}^2)$.

Supposons que vous voulez calculer la moyenne d'âge des habitants d'un pays et que vous tirez aléatoirement 1000 échantillons différents et calculez chaque fois la moyenne. Les moyennes calculées vont se répartir sur une courbe en forme de cloche, une courbe de Gauss. Nous y reviendrons lorsque nous parlerons des intervalles de confiance.

5.3.4 Vérifier la normalité des données et transformations

On peut de façon qualitative considérer le Q-Q plot (Quantile-Quantile) qui est un graphique qui compare les quantiles théoriques (sachant qu'on connait la moyenne et la variance, on peut calculer les quantile d'une variable qui suivrait une loi normale ayant cette moyenne et cette variance) et les quantiles observés. Si les points de ce graphique sont plus ou moins alignés le long d'une droite alors on peut conclure que la variable suit une loi normale.

Un Q-Q plot à l'aspect suivant, en prenant l'exemple du revenu des ménages :



Sur cet exemple on constate très clairement que les revenus des foyers ne suivent pas une distribution normale, car les points (en rouge) s'éloignent de la droite à partir d'un certain niveau de revenus.

En fait **beaucoup de méthodes statistiques sont basés sur l'hypothèse de la normalité des données**. Cette hypothèse mène à des tests qui sont simples et puissants comparés à certains autres tests n'étant pas basés sur cette hypothèse. Malheureusement, beaucoup de jeux de données n'ont pas une distribution normale.

Pour information : Outre une vérification visuelle sur la base d'un Q-Q-plot, il est également possible de réaliser des tests. Un des tests les plus courants et les plus anciens et le test de Kolmogorov-Smirnov.

S'il faut vraiment avoir une distribution normale pour pouvoir utiliser une méthode statistique, on peut essayer de transformer une variable qui ne suit pas cette loi. Il y a beaucoup de possibilités de transformations :

- racine carrée,

- logarithme naturel ln(x)

- transformations de Box-Cox ($\frac{X^{\lambda}-1}{\lambda}$)

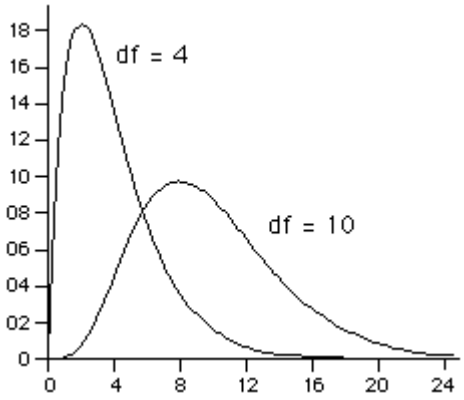
- etc.

5.4 Distribution du t et distribution du χ^2 (chi-carré)

Si plusieurs variables $Z_1, Z_2, ..., Z_n$ suivent une loi normale centrée réduite, la variable

V = $Z_1^2 + Z_2^2 + ... + Z_n^2$

suit une distribution du chi-carré. Cette distribution est asymétrique à droite si n est petit, mais plus n est grand et plus la distribution devient symétrique et en forme de cloche, et elle converge vers la distribution de la loi normale.



Dans ce graphique df signifie « degree of freedom » - degré de liberté. Dans le cas du test d'indépendance du chi-carré, il s'agit de (l-1)*(c-1).

Une variable qui suit une distribution du t peut s'obtenir en divisant une variable qui suit une loi normale centrée réduite N(0,1) par une variable qui suit une loi du chi-carré, qui a été divisée par son nombre de degrés de liberté, p.ex. :

$$t = \frac{Z}{\sqrt{V/n}} \sim t(n).$$

Cette distribution est symétrique autour de 0, mais est tout de même un peu différente de N(0,1) ; cela étant, lorsque n augmente elle converge vers la loi normale centrée réduite.

Nous retrouverons ces deux distributions lorsque nous étudierons le test d'indépendance du chi-carré et les régressions.

UNIVERSITÉ DE NEUCHÂTEL

TIRER DES CONCLUSIONS SUR LA BASE DE L'ECHANTILLON : LA STATISTIQUE INFERENTIELLE

6. INTERVALLES DE CONFIANCE

6. 1 Intervalle de confiance pour la moyenne

Selon le théorème central limite, Si n > 30 alors les moyennes calculées suivent approximativement une loi normale, donc $\bar{x} \sim N(\mu_{\bar{x}}, \sigma_{\bar{x}}^2)$, même si on ne connaît pas la distribution exacte des x_i.

Ceci nous permet de calculer le **risque d'erreur lié à l'échantillonnage**. Notre but n'est pas seulement de connaître la moyenne dans l'échantillon mais aussi de *tirer des conclusions sur la population de référence.*

On va donc chercher à calculer un intervalle de confiance tel que la probabilité que cet intervalle couvre la vraie valeur du paramètre de la population s'élève à 95%, cet I.C étant de la forme [$\bar{x} \pm e$], $\pm$ e étant *la marge d'erreur*. Car en fait, ce qui nous intéresse, ce n'est pas la moyenne dans l'échantillon, mais la moyenne dans toute la population.

On trouve notamment ces indications dans les sondages d'opinions : si vous y prêtez attention, il y a en général un $\pm e$ % mentionné dans l'article. Notons que cela présuppose que la façon dont a été conduit le sondage respecte les lois de la statistique inférentielle (basée sur le tirage d'un échantillon aléatoire) ; c'est loin d'être toujours le cas.

Posons α le risque de se tromper, en règle générale 5% dans les sciences sociales (au sens large). Mais on peut se montrer plus exigeant selon l'importance de la précision de l'I.C., p.ex. 99%.

On écrit: $P(\mu - e \leq \bar{x} \leq \mu + e) = 0.95 = 1 - \alpha$

On va recourir à la loi normale centrée réduite pour faire ce calcul. Il faut chercher a et b tels que $P(Z \leq b) = 97.5\%$ et $P(Z \leq a) = 2.5\%$. En regardant dans les tables, on trouve que :

$\Rightarrow$ a = -1,96 et b = 1.96

Posons $\sigma_{\bar{x}}$ = l'écart-type de la distribution des moyennes d'échantillon.

$$P(-1{,}96 \leq \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \leq 1{,}96) = 0{,}95, \text{ avec } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

L' I.C. pour la moyenne arithmétique vaut :

$$[\ \bar{x} - 1{,}96\ \frac{S}{\sqrt{n}}\ ;\ \bar{x} + 1{,}96 \frac{S}{\sqrt{n}}\ ]$$

On voit que plus la taille de l'échantillon est grande, plus l'IC est petit. Et surtout que l'IC **ne dépend pas de la taille de la population de référence.**

Prudence : n ne désigne pas toujours la taille de l'échantillon dans son entier, il peut s'agir d'un sous-échantillon sur lequel on estime un paramètre. Donc plus le sous-groupe de l'échantillon sur lequel est calculée la moyenne est petit, plus l'IC est large.

Rappelons aussi que le problème de constitution de l'échantillon reste entier même si on a une faible marge d'erreur : la base de sondage utilisée, le taux de non-réponse, les biais dans la collecte des données, etc.

6.2 Intervalle de confiance pour une proportion

La formule de l'IC avec un niveau de confiance de 95% (probabilité de couvrir la valeur) pour une proportion vaut donc (π est la proportion qu'on cherche à estimer):

$$p - 1{,}96\sqrt{\frac{s^2}{n}} \leq \pi \leq p + 1{,}96\sqrt{\frac{s^2}{n}}$$

Avec $\mathbf{s^2 = p(1\text{-}p)}$ – la variance d'une variable qui suit une distribution de Bernoulli - et n la taille de l'échantillon ou du sous-groupe de l'échantillon considéré, donc l'IC peut s'écrire :

$p \pm 1.96 \cdot \sqrt{\frac{p \cdot (100 - p)}{n}}$, si p est exprimé en %,

et la *marge d'erreur maximale* vaut : $\pm 1.96 \cdot \sqrt{\frac{2500}{n}}$

C'est la marge d'erreur maximale qui est généralement indiquée dans les sondages d'opinion.

Notons toutefois que les formules évoquées ci-dessus ne fonctionnent que pour les cas où l'échantillon n'excède pas 5% de la taille de la population de référence, sinon il faut recourir à un facteur correctif $\frac{N-n}{N-1}$ qu'on ajoute sous la racine carrée.

Deux remarques fondamentales :

i) cette formule ne s'applique que pour n≥30

ii) cette formule n'est valable que pour des échantillons aléatoires simples.

Il faut savoir que le plan de sondage (stratifié, par grappes, etc.) et la pondération peuvent faire varier la taille de l'I.C. Certains logiciels permettent de prendre en compte ces deux facteurs (p.ex. SPSS). Notons encore que, lorsqu'on compare deux proportions, si leur I.C. ne se recouvrent pas, il est très

probable que ces deux proportions soient significativement différentes l'une de l'autre. Dans le cas contraire, rien n'est moins sûr.

7. VARIABLES CATÉGORIELLES ET TABLEAUX DE CONTINGENCE

7.1. Généralités

Il s'agit d'un aspect particulièrement important de la statistique appliquée aux sciences humaines. En effet, les variables qualitatives jouent un rôle important, et les tableaux de contingence sont très fréquents dans les publications.

Abordons tout d'abord un point fondamental : l'interprétation de tableaux croisés. Cela paraît évident, malheureusement des fautes sont commises très fréquemment. Quand on lit des pourcentages dans un tableau croisé, il faut d'abord se demander: s'agit-il de pourcents en ligne, en colonne ou sur le total ?

7.2. Tableaux de contingence

Un tableau de contingence croisant deux variables A et B est un tableau qui a la forme suivante :

	1	**2**		**j**		**c**	**total**
1	n_{11}	n_{12}		n_{1j}		n_{1c}	n_{1+}
2	n_{21}						n_{2+}
i	n_{i1}			n_{ij}			n_{i+}
l	n_{l1}					n_{lc}	n_{l+}
total	n_{+1}	n_{+2}		n_{+j}		n_{+c}	n

n_{ij} est l'élément de la i^{ème} ligne et de la j^{ème} colonne.

n_{i+} est le nombre d'éléments sur la i^{ème} ligne = $\sum_{j=1}^{c} n_{ij}$

n_{+j} est le nombre d'éléments dans la j^{ème} colonne = $\sum_{i=1}^{l} n_{ij}$

Si les deux variables sont indépendantes, on sait par les lois des probabilités que
$P(A \cap B)=P(A)*P(B)$, donc :

P(ligne=i et colonne=j)=P(ligne=i)*P(colonne=j) $\Rightarrow \frac{n_{ij}}{n} = \frac{n_{i+}}{n} \cdot \frac{n_{+j}}{n} \Rightarrow n_{ij} = \frac{n_{i+} \cdot n_{+j}}{n}$

Donc la valeur attendue dans chaque case, si les deux variables sont indépendantes l'une de l'autre, vaut $\frac{n_{i+} \cdot n_{+j}}{n}$,

soit le nombre de personnes de toute la i^{ème} ligne multiplié par le nombre de personnes de toute la j^{ème} colonne, divisé par le nombre total d'individus.

7.3. Sur-représentation et sous-représentation

La première façon de mesurer ces écarts entre les effectifs théoriques et les effectifs observés, c'est de calculer le ratio :

Effectifs observés de l'évènement conjoint / Effectifs théoriques sous l'hypothèse d'indépendance.

La règle est simple : si ce rapport est >1, cela veut dire que le numérateur est plus grand que le dénominateur, ce qui veut dire que le groupe observé est **sur-représenté.** Si ce rapport <1, alors le numérateur est plus petit que le dénominateur et le groupe observé est **sous-représenté.**

On utilise généralement plutôt les fréquences que les nombres absolus, mais le principe est le même. Pour illustrer cela graphiquement, on peut p.ex. marquer en gris les croisements sur-représentés et utiliser une couleur de plus en plus foncée à mesure que la sur-représentation augmente, par exemple :

1< X <1.5

1.5 < X < 2

2 < X < 3

X > 3

Appliquons cela à un tableau qui compare les niveaux de formation au sein des couples :

UNIVERSITÉ DE NEUCHÂTEL

Niveaux de formation croisés du chef de ménage et du partenaire

	Niveau de formation du partenaire								
Niveau de formation du chef de ménage	Non actif	Scolarité obligatoire	Formation prof.	Maturité	Form. prof. sup.	Ecole prof. sup.	Université	autres form.	Total
Scolarité obligatoire									
nombre	277	241	93	6	5	0	1	3	**626**
%	*9.22*	*38.38*	*5.12*	*2.76*	*2.48*	*0.00*	*0.58*	*6.25*	***10.14***
Formation prof.									
nombre	1542	288	1227	66	48	25	16	22	**3234**
%	*51.35*	*45.86*	*67.49*	*30.41*	*23.76*	*30.12*	*9.25*	*45.83*	***52.40***
Maturité									
nombre	126	14	60	32	4	8	9	3	**256**
%	*4.20*	*2.23*	*3.30*	*14.75*	*1.98*	*9.64*	*5.20*	*6.25*	***4.15***
Form. prof. Sup.									
nombre	389	42	229	26	75	9	9	10	**789**
%	*12.95*	*6.69*	*12.60*	*11.98*	*37.13*	*10.84*	*5.20*	*20.83*	***12.78***
Ecole prof. Sup.									
nombre	255	21	102	29	23	18	16	3	**467**
%	*8.49*	*3.34*	*5.61*	*13.36*	*11.39*	*21.69*	*9.25*	*6.25*	***7.57***
Université									
nombre	357	11	74	56	45	21	121	0	**685**
%	*11.89*	*1.75*	*4.07*	*25.81*	*22.28*	*25.30*	*69.94*	*0.00*	***11.10***
Autres formations									
nombre	57	11	33	2	2	2	1	7	**115**
%	*1.90*	*1.75*	*1.82*	*0.92*	*0.99*	*2.41*	*0.58*	*14.58*	***1.86***
Total									
nombre	**3003**	**628**	**1818**	**217**	**202**	**83**	**173**	**48**	**6172**
%	***100***	***100***	***100***	***100***	***100***	***100***	***100***	***100***	***100***

Sources : OFS, Enquête sur les revenus et la consommation des ménages, en 1990, échantillons mensuels non pondérés

On a ainsi une illustration très parlante d'un phénomène bien connu en sciences sociales : l'endogamie sociale (ou homogamie de classe), c.à.d. le fait qu'au sein des couples, les conjoints tendent à avoir des niveaux socio-économiques similaires.

UNIVERSITÉ DE NEUCHÂTEL

8. LES TESTS STATISTIQUES ET LES MESURES D'ASSOCIATIONS

8.1. Les tests en statistique inférentielle

Toute démarche scientifique repose sur la formulation d'une hypothèse sur la population, suivie d'une collecte de données puis du rejet ou du non-rejet de l'hypothèse sur la base de l'écart existant entre les données observées et les valeurs prédites par l'hypothèse. Ces tests jouent notamment un rôle fondamental dans les approches expérimentales, car on travaille avec des groupes de sujets assez limités en nombre, et on veut donc savoir si les différences observées sont statistiquement significatives.

Nous avons déjà parlé des estimations (ponctuelles ou par intervalle de confiance) : il s'agissait d'une démarche de quantification. Dans le cas d'un test d'hypothèse, il s'agit d'une **validation**.

On formule une hypothèse de départ, qu'on appelle *hypothèse nulle* et qu'on note H_0. La conclusion du test contient un risque d'erreur. En fait il y en a deux :

α : Probabilité de rejeter l'hypothèse H_0 alors qu'elle est vraie, c'est l'erreur de 1ère espèce

β : Probabilité d'accepter H_0 alors qu'elle est fausse, c'est l'erreur de 2ème espèce.

Les deux autres cas de figures sont bien évidemment ceux qu'on recherche : ne pas rejeter H_0 quand elle est vraie et rejeter H_0 quand elle est fausse.

La marche à suivre est la suivante :

i) formuler H_0 et définir son alternative H_1
ii) définir α, en général 5%
iii) définir un critère qui nous permet de tirer une conclusion : c'est la statistique du test T dont on devra connaître la distribution (N, t, χ^2, F, etc.)
iv) définir une région de rejet telle que si la statistique du test est plus grande (ou plus petite) qu'une valeur critique, alors l'hypothèse est rejetée.
v) prélever un échantillon
vi) calculer la statistique du test sur l'échantillon et tirer une conclusion.

Il existe de nombreux tests en statistique inférentielle : des tests d'indépendance, des tests d'adéquation, des tests sur les moyennes, sur la variance, sur les coefficients de régression, etc.

Une autre approche est possible, qui est celle dite de la « p-valeur » ou valeur p. Plutôt que de définir une zone de rejet, on calcule la probabilité de trouver une valeur encore plus « extrême » que la statistique du test calculée. Si cette probabilité est petite, on rejette l'hypothèse, c'est-à-dire si la valeur p est $< \alpha$, qui vaut généralement 0,05 ou 0,01. C'est l'information qu'on obtient en général dans les tableaux statistiques générés par les logiciels habituels.

Il faut souligner que la valeur p **n'est pas** :

- une indication de l'importance du phénomène qu'on étudie, elle dit seulement s'il y a ou non un effet
- la probabilité que l'hypothèse soit vraie.

8.2. Le test d'indépendance du χ^2

Jusqu'à présent, nous avons analysé de façon qualitative la différence existant entre les effectifs observés et les effectifs attendus. Ici il s'agit d'une approche plus précise, basée sur la statistique inférentielle et les tests.

Nous avons vu que le critère de décision sur l'éventuelle indépendance de ces deux variables dépend de la différence entre les fréquences observées et les fréquences attendues. C'est en comparant ces deux grandeurs que nous avons déterminé quelle couleur attribuer à une case.

Il nous faut déterminer une mesure globale des écarts entre les effectifs observés et les effectifs attendus, en utilisant les notations suivantes :

n_{ij} pour les fréquences observées

n^*_{ij} pour les fréquences attendues sous l'hypothèse d'indépendance (ou théoriques)

T est la statistique du test, qu'on calcule de la manière suivante :

$$T = \sum_{i=1}^{c} \sum_{j=1}^{l} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

C'est-à-dire la somme de (effectifs observés – effectifs attendus)^2 / effectifs attendus,
pour chaque case du tableau de contingence.

On sait que T suit une distribution χ^2, on note T ~ χ^2.

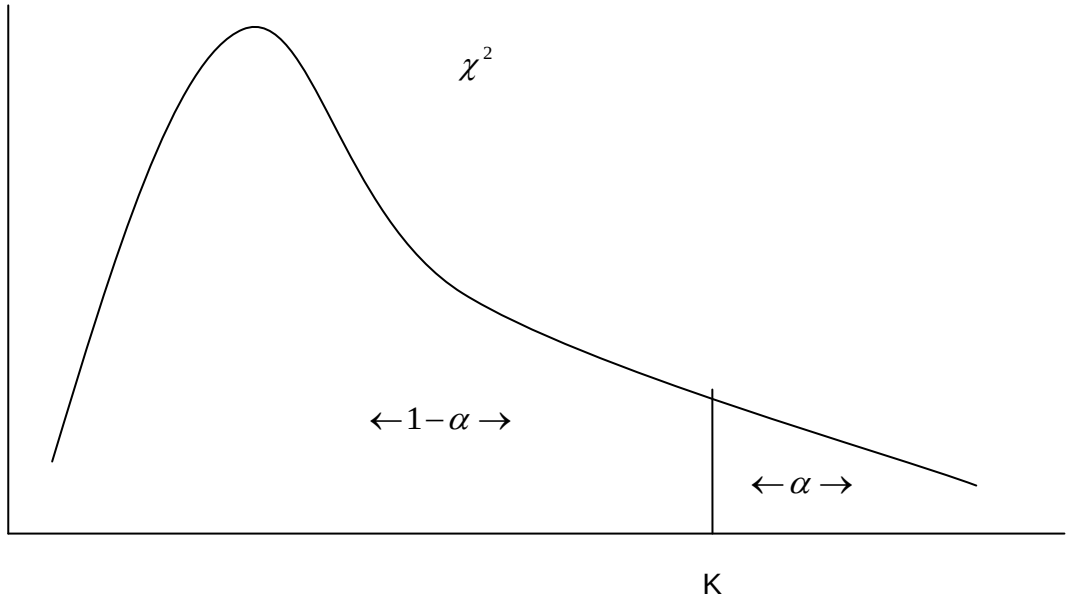
Calculer la probabilité que T soit plus petit qu'une valeur seuil est assez compliqué, mais on dispose de tables, comme pour la loi normale centrée réduite.

On rejette l'hypothèse d'indépendance H_0 si T > $\chi^2_{(l-1)\cdot(c-1),\alpha}$

pour l le nombre de lignes dans le tableau, et c le nombre de cases.

Cela dépend donc, entre autres, du nombre de lignes et de colonnes du tableau de contingence, ce qui est logique puisque plus il y a de cases dans le tableau, plus le nombre de différences possibles est élevé. Ce calcul dépend également de l'erreur de première espèce (α), bien entendu.

Représentons cela graphiquement :



Remarque *très* importante :

Pour que le test soit valable, on ne doit pas avoir plus de 20% des cellules avec un effectif attendu inférieur à 5, et aucune avec un effectif observé nul. Si ces conditions ne sont pas remplies, on peut regrouper certaines modalités.

8.3. Les mesures d'associations

On utilise généralement les grandeurs suivantes, en posant T la statistique du test et n le nombre de répondants :

Le coefficient de Cramer : V = $\sqrt{\frac{T}{n \cdot \min(c-1, l-1)}}$

Un cas particulier du coefficient de Cramer est le coefficient $\Phi = \sqrt{\frac{T}{n}}$,

c'est le coefficient de Cramer pour une table de contingence 2x2.

On utilise aussi le coefficient de contingence C = $\sqrt{\frac{T}{T+n}}$

Ces mesures prennent des valeurs **entre 0 et 1**. Plus la valeur est élevée, plus le lien est fort.

En fait il y a deux types de réponses qu'on apporte :

- la relation est-elle forte ? → mesures d'association
- y a-t-il une relation significative ? → test d'indépendance

8.4. Le test du t sur deux moyennes

Il est fréquent de vouloir comparer la moyenne d'un groupe A à celle d'un groupe B. Ceci est particulièrement le cas dans les approches expérimentales, p.ex. de vouloir comparer la moyenne obtenue par divers groupes de sujets. Par exemple, on veut comparer le temps de réaction moyen d'un groupe expérimental qui a été soumis à une condition particulière à celui d'un groupe dit de contrôle n'ayant pas été soumis à cette condition. On veut savoir si la différence observée – car il est quasiment impossible d'observer exactement la même moyenne, même si la condition expérimentale n'a aucun effet – est une pure coïncidence ou si elle est, au contraire, statistiquement significative.

Afin de pour pouvoir « trancher » de manière univoque, on peut recourir à un test nous permettant de savoir si l'écart observé entre ces deux moyennes (deux groupes de la population ou entre le groupe expérimental et le groupe de contrôle dans une étude expérimentale) est statistiquement significatif.

L'hypothèse à vérifier, ainsi que l'hypothèse alternative, sont les suivantes :

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$ (ce qu'on appelle un test bilatéral).

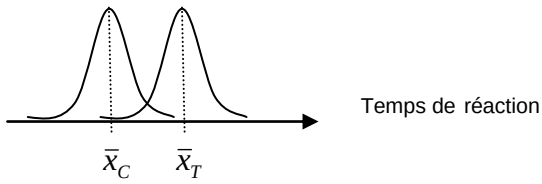
La statistique du test vaut, si les deux échantillons sont de taille et de variance différente :

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

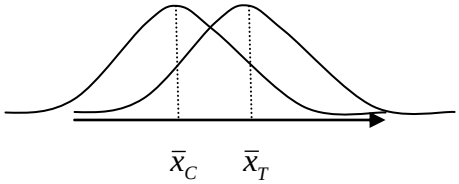
Avec n_1, n_2 les deux sous-échantillons sur lesquels on a calculé $\bar{x}_1$ et $\bar{x}_2$, les deux barres verticales symbolisant la valeur absolue, donc le numérateur est toujours positif. σ_1^2 et σ_2^2 sont les variances calculées pour les deux sous-échantillons. Cette formule est valable pour deux (sous-) échantillons indépendants.

En effet, ce n'est pas seulement la différence entre les deux moyennes qui est importante, mais également la dispersion au sein de chaque groupe (donc la variance). Comparons les deux situations expérimentales suivantes :

Expérience 1 (T désigne le groupe traitement et C le groupe de contrôle):



Expérience 2 :



Dans les deux cas, la distance entre la moyenne du groupe traitement et celle du groupe contrôle est identique, pourtant on voit bien que la situation n'est vraiment pas la même. En effet, dans le 1er cas, les 2 groupes sont clairement distincts, alors que dans le second, les deux groupes se recoupent largement et beaucoup de sujets dans les deux groupes ont des temps de réaction similaires. D'où la nécessité d'inclure les variances dans le calcul de la statistique du test (t).

On sait que la statistique du test décrite plus haut suit une distribution t de Student, et que la valeur critique est la suivante : $t_{n_1+n_2-2,\frac{\alpha}{2}}$

Il se trouve que la distribution t, à partir d'un certain nombre de degrés de liberté, c'est-à-dire à partir d'une certaine taille d'échantillon, ressemble beaucoup à une loi normale centrée réduite. Pour être plus précis, on peut chercher la valeur critique dans une table de la loi normale à partir de $n_1 = 30$ et $n_2 = 30$ (ou $n_1 + n_2 > 40$). Si on travaille sur une enquête réalisée auprès d'un échantillon représentatif, on a souvent beaucoup plus de cas que cela ; ce n'est par contre pas toujours le cas dans les expériences de laboratoire.

A partir d'un nombre suffisant d'observations, on peut dire que le seuil pour $\alpha = 0{,}05$ est :

$Z_{\alpha/2}$ de la loi normale centrée réduite, c'est-à-dire $Z_{0.75}$, qui vaut **1,96**.

Si l'on fixe $\alpha = 0{,}01$, cette valeur critique devient 2,58.

Donc si la statistique ci-dessus est supérieure à 1,96 (ou 2,58), si les échantillons ont une taille suffisante, on rejette l'hypothèse nulle, donc il y a une différence significative entre les deux moyennes. Par contre, dans le cadre d'une approche expérimentale, il se peut que les groupes soient plus petits que les limites mentionnées ci-dessus. Dans ce cas, les valeurs critiques sont à chercher dans une table de la distribution du t de Student.

Remarques importantes :

Il faut souligner ici que les **échantillons** sont **indépendants**. Si deux échantillons sont liés (p.ex. deux mesures sont réalisées sur la même population), le test est un peu différent.

D'autres part, pour que le test soit valide, il y a des hypothèses techniques « ennuyeuses » : les **données doivent suivre une loi normale et l'écart-type doit être le même dans les deux échantillons**. Donc avant de réaliser un test du t, il faut s'assurer que la distribution des résultats a plus ou moins une forme de cloche, et que les variances ne sont pas trop différentes.

Si l'on doit comparer les moyennes de trois groupes ou plus, alors on procède à une analyse de variance (cf. point 9 ci-dessous).

8.5. Autres mesures d'associations et autres tests

Après avoir parlé des mesures d'associations pour des variables nominales (coefficient de Cramer et de contingence), il nous faut encore signaler deux coefficients mesurant **le lien entre deux variables ordinales** (par exemple le lien entre le niveau de formation des deux conjoints, ou le lien entre niveau de formation et position hiérarchique dans l'entreprise) :

1) le coefficient de corrélation des rangs ρ **de Spearman** : on observe le « classement » d'un individu sur deux dimensions et on regarde si les rangs sont liés:

$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$, avec d la différence entre les rangs x_i - y_i et n le nombre d'observations, si les ex aequo représentent plus de 20% des rangs, alors il faut modifier la formule.

2) le coefficient τ **de Kendall**, qui compare également les rangs : on compare deux par deux les individus i et j ; si $x_i - x_j$ et $y_i - y_j$ ont le même signe, la paire est concordante, et si ce n'est pas le cas elle est discordante. On compte le nombre de chaque occurrence (concordant et discordant), n_c et n_d et on calcule ce coefficient : $\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$. S'il y a des ex aequo, une autre formule s'applique.

Les logiciels statistiques proposent ces deux coefficients, notamment SPSS. Ils varient **entre -1 et 1**, 1 étant la correspondance parfaite, -1 également mais dans le cas où deux variables varient en « sens inverse », c.-à-d plus on a un rang élevé sur une dimension et plus on a un rang bas sur l'autre dimension. Le logiciel fournit aussi un test qui indique si le coefficient est significativement différent de 0, donc s'il y a un lien significatif entre les deux variables (il indique la p-valeur : si elle est plus petite que 0,05, le lien est significatif).

Plus loin nous parlerons des coefficients de corrélation de Pearson qui mesurent le lien entre deux variables quantitatives.

Dans le cas où on mesure le lien entre **une variable ordinale et une variable nominale**, on peut réaliser des tests : soit le test de Mann-Whitney, soit le test de Kruskal-Wallis. La variable nominale détermine l'appartenance à un groupe, p.ex. femmes et hommes, ou catholiques-protestants-musulmans-autres confessions. Le test vous dit s'il y a une différence significative entre les groupes

pour la variable ordinale, p.ex. la réponse à un sondage d'opinion (tout à fait d'accord, plutôt d'accord, pas vraiment d'accord, pas d'accord du tout). Là aussi, le logiciel vous indique la p-valeur, et si elle est plus petite que 0,05, la différence entre les groupes est significative.

9. ANALYSE DE VARIANCE (ANOVA)

L'analyse de variance permet de comparer la moyenne d'une **variable quantitative** observée dans divers groupes ; l'appartenance à un groupe, est, elle, une **variable nominale**. En d'autres termes, on « explique » une variable quantitative par une variable nominale (l'appartenance à un groupe). Cette technique est dérivée de l'approche expérimentale en sciences humaines, notamment en psychologie, dont l'idée est de comparer plusieurs groupes étant soumis à des conditions expérimentales différentes pour voir si ces conditions induisent des différences statistiquement significatives. Notons ici aussi que le test ne démontre pas qu'il y a un lien de causalité, il montre que la moyenne est significativement différente entre les groupes.

9.1. Hypothèses

Pour pouvoir effectuer une ANOVA, il faut vérifier que certaines conditions sont remplies :

1. la distribution de la variable dépendante dans chaque groupe suit une distribution normale,

2. l'homogénéité des variances (ou homoscédasticité), autrement dit, il faut que la variance de la variable quantitative soit la même dans chaque groupe. On peut vérifier cela au moyen du test de Levene, proposé dans le logiciel SPSS,

3. que les échantillons soient extraits de la même population de référence de manière aléatoire et qu'ils soient indépendants les uns des autres, ou sélectionnés de façons indépendante dans k groupes différents. La violation de cette hypothèse pose problème ; il existe des techniques ANOVA pour les mesures répétées, mais nous ne les verrons pas ici.

9.2. Déroulement du test

La marche à suivre est conforme à celle plus générale de la réalisation de tests statistiques:

1. formuler les hypothèses $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ (les moyennes sont égales)
   et H_1: les k moyennes ne sont pas toutes égales.

2. réaliser une série de calculs afin de comparer des variances, d'où le nom de la méthode, en calculant la statistique du test qui suit une distribution $F_{k-1,n-k,\alpha}$, avec k le nombre de groupes et n le nombre total d'observations.

3. ensuite la statistique du test est comparée, comme dans tout test statistique, à une valeur critique Fc; si F>Fc, on rejette l'hypothèse H_0, il y a dans ce cas un impact significatif de la variable indépendante (l'appartenance à un groupe) sur la variable qu'on cherche à expliquer.

On peut réaliser une ANOVA avec une variable « dépendante » dichotomique (prenant les valeurs 0 et 1); en outre, l'ANOVA semble peu sensible à la non-normalité des distributions si la taille des groupes

est suffisamment grande (plus de 20 cas) et si la taille des groupes est assez similaire; il faut toutefois s'abstenir si la distribution est très asymétrique.

Plusieurs approches existent:

a) ANOVA simple: l'appartenance aux groupes est définie par une seule variable, et on ne mesure qu'une variable quantitative.

b) On peut imaginer un cas avec une variable qualitative mais plusieurs mesures quantitatives (MANOVA).

c) On peut également avoir plusieurs variables qualitatives et une variable quantitative (*factorial ANOVA*).

d) On peut également avoir plusieurs variables quantitatives et qualitatives.

Nous ne traitons ici que du premier cas de figure (ANOVA simple), par exemple une recherche sur les temps de réaction dans deux groupes expérimentaux ne différant que par une variable. L'idée fondamentale est de comparer d'une part la variabilité observée entre les groupes expérimentaux, et d'autre part, la dispersion au sein de chaque groupe expérimental.

Si les différences entre les groupes sont importantes et qu'au sein des groupes on observe une variation assez faible, alors on a de bonnes raisons de penser que la « variable indépendante » a un impact significatif. Cela dit cette méthode ne s'applique pas qu'aux situations expérimentales. On peut par exemple appliquer cette méthode en prenant comme variable qualitative le fait de vivre dans une région, et on peut ensuite voir s'il y a des différences significatives entre régions pour des variables quantitatives.

Donc, la statistique du test va mettre en lien la variabilité *intergroupe* et la variabilité *intragroupe*, et pour cela on va se focaliser sur les écarts à la moyenne. D'une part les écarts à la moyenne de l'ensemble des participants, et d'autre part les écarts à la moyenne au sein de chaque groupe. Remarque sur la notation : en français les termes utilisés sont intergroupe et intragroupe, ce qui ne facilite pas le choix d'un indice ; on prend donc les indices en anglais (B=between=intergroupe et W=within=intragroupe).

On calcule la somme des carrés intragroupe :

$SC_W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i.})^2$, avec k= le nombre de groupes (de conditions expérimentales), n_i le nombre d'individus dans le i^{ème} groupe, $\overline{Y}_{i.}$ la valeur moyenne dans le i^{ème} groupe, et Y_{ij} et la valeur de chaque individu.

La somme des carrés intergroupe vaut :

$SC_B = \sum_{i=1}^{k} n_i (\overline{Y}_{i.} - \overline{Y})^2$, avec $\overline{Y}$ la moyenne de toutes les observations.

Et la somme des carrés totale $SC_T = SC_B + SC_W$.

La statistique du test est définie comme le ratio de $\frac{SC_B}{k-1}$, k = nombre de conditions expérimentales (nombre de groupes) et de $\frac{SC_W}{n-k}$, n le nombre total de participants.

La statistique du test se calcule de la façon suivante :

$$T = \frac{SC_B / k-1}{SC_W / n-k} = \frac{SC_B}{SC_W}\frac{n-k}{k-1},$$

et on rejette H_0 si T > $F_{k-1,n-k,\alpha}$, dont on trouve les valeurs dans les tables appropriées.

Rejeter H_0 =la « variable indépendante » n'a aucun effet sur la « variable dépendante » est une chose, mais cela n'indique pas si cet impact est fort ou faible.

Pour cela on peut calculer le ratio $\frac{SC_B}{SC_T}$, qu'on note R^2 et qui varie entre 0 et 1 : c'est le % de la variance de la variable dépendante qui est expliquée par l'appartenance aux groupes ; dit autrement, plus R^2 est élevé, plus le « pouvoir explicatif » de la « variable indépendante » est élevé.

En général, les explications en sciences sociales et en psychologie reposent sur plus d'une variable indépendante. On peut réaliser des ANOVA avec 2 « conditions expérimentales », voire plus. Prenons p.ex. une ANOVA avec 2 variables indépendantes, x ayant 2 modalités et y 3 modalités, alors cela crée 6 conditions expérimentales différentes, qui correspondent à un tableau croisé de 2 lignes et 3 colonnes. Ici on a un exemple avec le même nombre d'observations par cellule (20) :

	Var 2, modalité 1	Var 2, modalité 2	Var 2, modalité 3
Var 1, modalité 1	20 observations	20 observations	20 observations
Var 1, modalité 2	20 observations	20 observations	20 observations

Si le nombre d'observations est le même dans chaque cellule alors l'équation devient :

$$SC_T = SC_L + SC_C + SC_{LC} + SC_W =$$

$$\sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{20}(\bar{Y}_{i..} - \bar{Y}_{...})^2 + \sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{20}(\bar{Y}_{.j.} - \bar{Y}_{...})^2 + \sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{20}(\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} - \bar{Y}_{...})^2 + \sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{20}(Y_{ijk} - \bar{Y}_{ij.})^2$$

$\bar{Y}_{i..}$ est la moyenne de la ième ligne, $\bar{Y}_{.j.}$ la moyenne de la jème colonne, $\bar{Y}_{ij.}$ la moyenne de la cellule à la ième ligne et jème colonne et $\bar{Y}_{...}$ la moyenne de toutes les observations.

Comme on le constate, le calcul est rendu plus compliqué : il y a maintenant une somme des carrés par ligne (SC_L), une somme des carrés par colonne (SC_C), et un terme incluant des effets d'interaction (SC_{LC}). Le calcul est encore plus compliqué si l'on a des cellules qui n'ont pas le même effectif, ce qui est généralement le cas, sauf dans un contexte expérimental.

UNIVERSITÉ DE NEUCHÂTEL

10. CORRELATIONS ET REGRESSIONS

Nous retournons maintenant aux **variables quantitatives**. Nous nous intéressons aux liens linéaires qui peuvent unir deux variables de ce type.

10.1. Covariance et coefficient de corrélation de Pearson

Avant de passer au coefficient de Pearson r, nous devons introduire le concept de covariance.

La covariance mesure le lien qui unit deux variables en tenant compte des écarts à la moyenne de chaque variable :

Covariance de X et Y (empirique) = Cov (X,Y) = $\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$, le paramètre étant $E[(X-\mu_x)(Y-\mu_x)]$. Cette valeur peut varier entre $-\infty$ et $+\infty$.

Si on calcule Cov(X,X), on obtient $\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x}) = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$, qui est en fait la variance de X.

Le coefficient de corrélation de Pearson (ou Bravais-Pearson) vaut :

$$r = \frac{Cov(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}} = \frac{n\sum_{i=1}^{n} x_i \cdot y_i - (\sum_{i=1}^{n} x_i) \cdot (\sum_{i=1}^{n} y_i)}{\sqrt{(n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2) \bullet (n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2)}}$$

Le coefficient de Pearson varie **entre -1 et 1**.

Si les couples de points (x_i , y_i) sont à peu près alignés sur une droite alors r s'approche de 1 ou de -1.

Le coefficient est positif (>0) si y augmente quand x augmente et négatif (<0) si y diminue quand x augmente.

Lorsque r a une valeur proche de 0, on peut considérer que le lien est très faible voire inexistant *le long d'une droite.* Cela ne signifie pas nécessairement qu'il n'existe aucun lien entre les variables ! Il se peut qu'il existe un lien curvilinéaire. En outre il faut dire que, si les deux variables sont indépendantes, la corrélation sera nulle, mais que la réciproque n'est pas vraie, et on note :

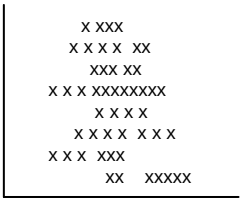
X, Y indépendantes $\Rightarrow$ corr(X,Y)=0

corr (X,Y) = 0 $\not\Rightarrow$ X, Y indépendantes.

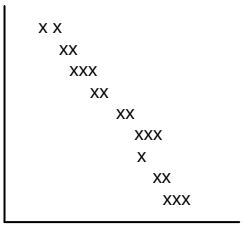
Par contre la contraposée est vraie :

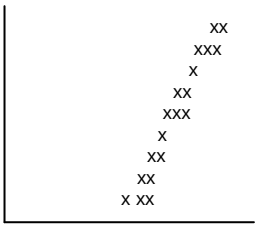
$\text{corr}(X,Y) \neq 0 \quad \Rightarrow$ X, Y ne sont pas indépendantes

Donnons quelques exemples graphiques, sous forme de *diagrammes de dispersion* :

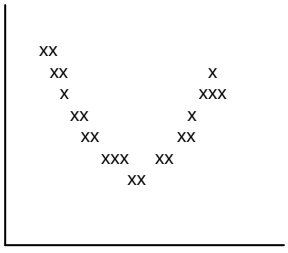


Dans ce cas, c.à.d un nuage de points dans lequel ne se dégage aucune tendance linéaire, on obtient une valeur du coefficient de Pearson $r \cong 0$





Dans ces deux cas, dans lesquels les points sont plus ou moins alignés, $r \cong -1$ resp. 1.



Dans ce cas aussi $r \cong 0$, pourtant on voit bien qu'il y a un lien entre les deux variables, mais celui-ci n'est pas linéaire.

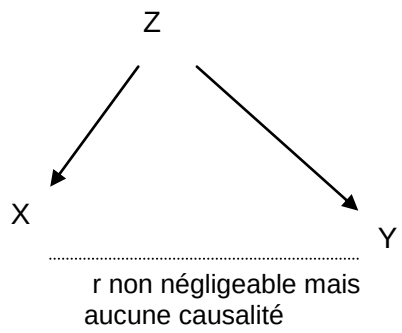
On peut avoir, entre deux variables, des liens quadratiques, logarithmiques, exponentiels, etc. Il faudrait toujours tracer le diagramme de dispersion pour s'en assurer.

Deux remarques s'imposent:

- Si le coefficient de corrélation est nul, cela ne signifie pas qu'il n'existe aucun lien entre les variables (cf. ci-dessus)

- Le coefficient de corrélation entre X et Y peut être assez élevé, ce qui indique une association de type linéaire entre les deux variables. Mais il ne faut pas l'interpréter automatiquement en termes de causalité ! Il se peut en fait que les deux variables n'aient, directement, rien à voir l'une avec l'autre mais dépendent toutes deux d'une troisième variable Z.

On peut illustrer cette dernière remarque de la façon suivante :



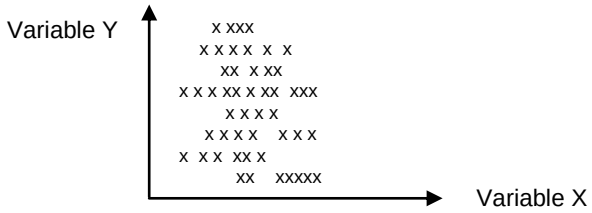
Pour vérifier que le coefficient est significativement différent de zéro, il faut réaliser un test ayant pour hypothèse nulle : r = 0 et l'hypothèse alternative $r \neq 0$.

La statistique du test vaut $t = r\sqrt{\frac{n-2}{1-r^2}}$, avec n le nombre d'observations.

Cette statistique suit une distribution t de Student et a n-2 degrés de liberté, n étant la taille de l'échantillon. Si cette statistique est supérieure à la valeur critique ($t_c \rightarrow 1{,}96$ pour n grand et $\alpha = 0{,}05$), on rejette l'hypothèse H_0 : r = 0.

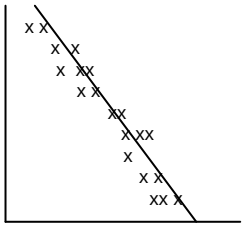
10.2. Régression linéaire simple

Nous avons déjà eu l'occasion de décrire des diagrammes de dispersion, qui ont l'apparence suivante :



Un tel diagramme permet de visualiser un ensemble de couples d'observations (x_i , y_i)

On souhaite calculer la droite qui s'approche le plus des points, c'est-à-dire qui minimise les distances entre les points et la droite; cette droite est **la droite de régression** :



Son équation peut s'écrire **y = a + bx.**

En fait y est une fonction de x : y est la variable dépendante (c.à.d la variable qu'on veut expliquer) et x est la variable indépendante ou explicative. Le terme **a** indique quelle est *l'ordonnée à l'origine*, c.à.d la valeur que prend y quand x vaut 0, et **b** est la *pente* de la droite.

Il faut donc calculer une droite telle que la somme des écarts de chaque point à cette droite soit la plus petite possible. En effet, la plupart des points ne sont pas sur la droite de régression. Il faut donc additionner les termes $y_i - (a + bx_i)$.

On recourt à la méthode des moindres carrés :

Q(a,b) = $\sum_{i=1}^{n} (y_i - (a + bx_i))^2$. On doit trouver le minimum de cette somme.

On élève les écarts à la droite de régression au carré pour que les distances des points se trouvant au-dessus de la droite (distances > 0) ne s'annulent pas avec les distances des points se trouvant en dessous de la droite (distances < 0).

On obtient les paramètres suivants (pour alléger la notation, on écrit $\sum$ pour $\sum_{i=1}^{n}$) :

$$\hat{a} = \bar{y} - \hat{b} \cdot \bar{x}$$

$$\hat{b} = \frac{n\sum x_i y_i - (\sum x_i \cdot \sum y_i)}{n\sum x_i^2 - (\sum x_i)^2} = r \cdot \sqrt{\frac{\sum (y_i - \bar{y})^2}{\sum (x_i - \bar{x})^2}}$$

Cette droite passe par les points (0, â) et ($\bar{x}, \bar{y}$).

Notons y_i la valeur observée de la variable Y et $\hat{y}_i$ la valeur prédite par la droite de régression.
Les résidus valent $e_i = y_i - \hat{y}_i$.

L'équation qu'on obtient est donc la suivante : y = a + bx + e.

On peut démontrer que la variation totale = variation au niveau des résidus + la variation expliquée par la régression.

Cela peut s'écrire :
$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

Somme des carrés totale (SCT) = Somme des carrés liés aux résidus, c.à.d les termes e (SCE) + somme des carrés liés à la régression (SCR).

Ces termes permettent de juger le « pouvoir explicatif » de la régression :

En effet le terme $\mathbf{R^2} = \frac{SCR}{SCT}$, qui varie entre 0 et 1,

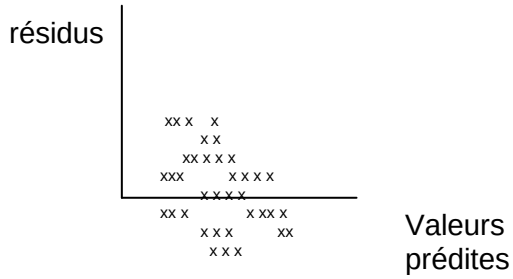
s'appelle le **coefficient de détermination** et s'interprète comme le pourcentage de la variabilité de y qui est liée à la variation de x (dans le modèle de régression). Dans le cas de la régression simple, $R^2 = r^2$.

Le but de la régression est double : d'une part il s'agit de résumer et modéliser le lien qui existe entre deux variables, et d'autre part il s'agit de prédire quelle valeur prendrait y pour un x qui n'a pas été observé. Notons que c'est l'aspect le plus délicat de la régression, c.à.d la prédiction en dehors des valeurs observées de x, car rien ne nous assure que la tendance linéaire observée se poursuit partout.

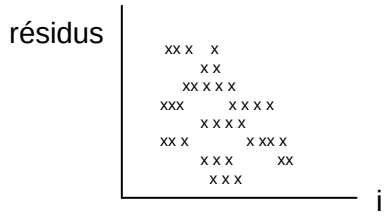
Il y a des **hypothèses techniques** à respecter qui concernent les résidus, c'est très important :

- Les résidus suivent une distribution normale de moyenne 0, $e_i \sim N(0, \sigma^2)$

- l'hypothèse de l'homoscédasticité, c'est-à-dire que $Var(\varepsilon) = \sigma^2$, c'est-à-dire que les erreurs n'augment pas (ou ne diminuent pas) quand la valeur prédite augmente, mais que les erreurs sont un peu « distribuées au hasard »

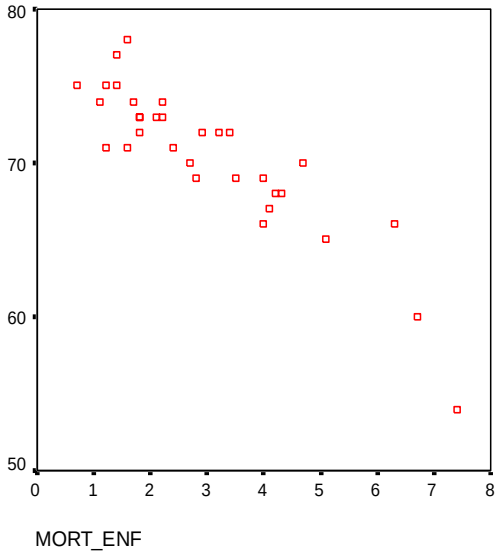


- en plus il faut vérifier que les termes d'erreur sont indépendants $E(\varepsilon_i \varepsilon_j)=0$, pour $i \neq j$, ce qui n'est pas le cas lorsqu'on utilise des séries temporelles p.ex.



Donnons un exemple :

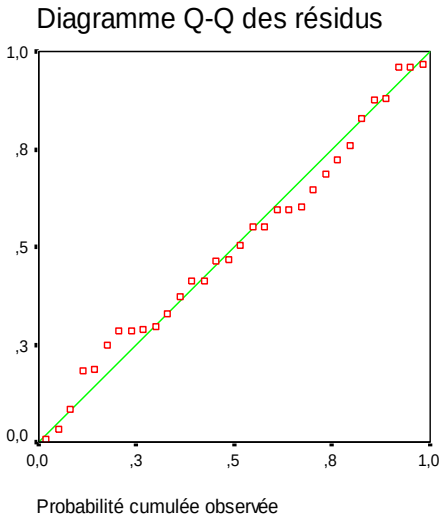
Un chercheur / une chercheuse souhaite expliquer l'espérance de vie par le taux de mortalité infantile en Amérique latine, centrale et dans les Caraïbes. On constate une tendance linéaire très marquée :



Calculons la droite de régression :
Espérance_vie = 78.002 – 2,514*mortalité_infantile

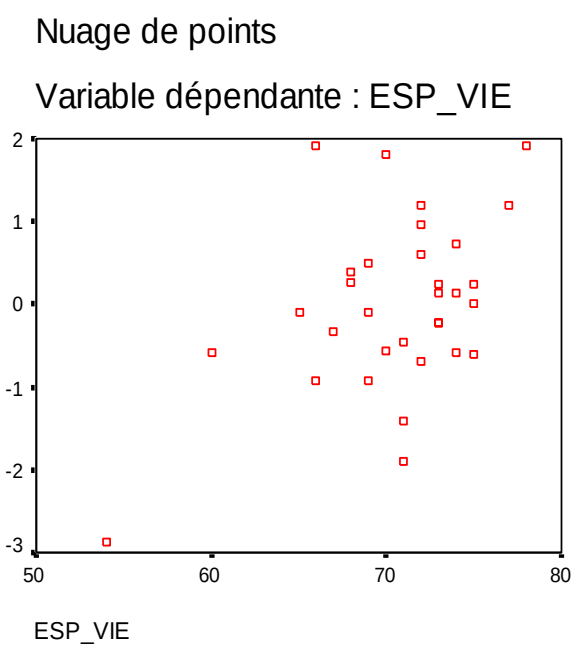
Vérifions certaines hypothèses techniques :

i) Normalité



Cela n'a pas l'air trop s'éloigner de la droite, donc on peut partir du principe les résidus suivent une distribution normale.

ii) Homoscédasticité :



Il n'y a pas un lien très évident entre les résidus et la prédiction, il y a une « bande » autour de 0, cette hypothèse technique semble être respectée.

iii) Indépendance des résidus :

On peut faire un diagramme de dispersion des couples de résidus (e_i, e_{i+1}), ou des couples (e_i, i).

On peut également calculer la statistique de Durbin-Watson. La statistique de ce test (D) est la suivante $\sum_{i=2}^{n} \frac{(e_i - e_{i-1})^2}{\sum_{j=1}^{n} e_j^2}$

et l'hypothèse d'indépendance n'est pas rejetée si D ≈ 2. Dans notre cas D= 2.482, ce n'est pas trop mal.

Partons donc du principe que les hypothèses techniques sont vérifiées. En réalité, on pourrait être plus rigoureux (il existe des tests pour la normalité et l'homoscédasticité).

Notons encore que R^2 = 0,803, donc environ 80% de la variabilité des espérances de vie est expliquée par la mortalité infantile (donc par ce modèle de régression).

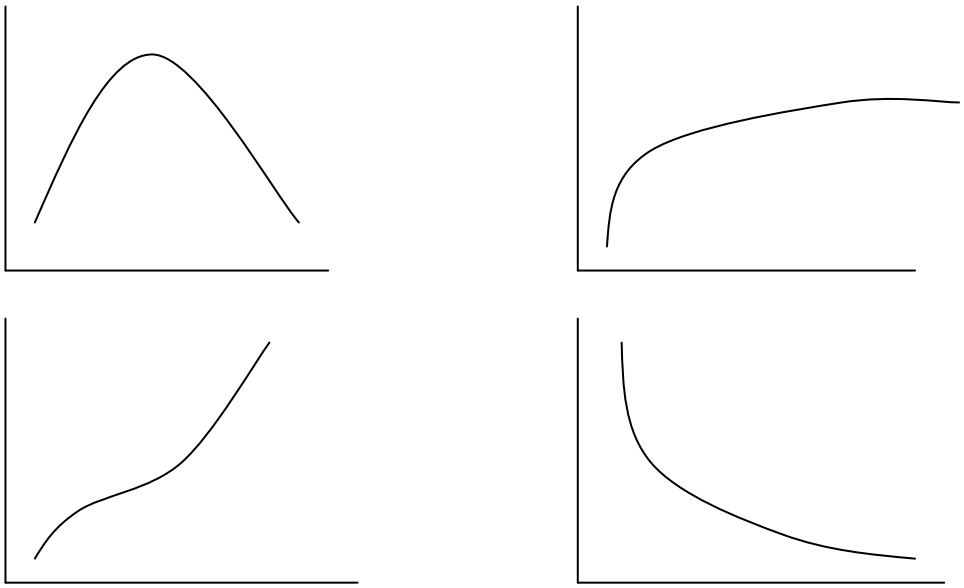
10.3. Régressions linéaires dans les paramètres et/ou linéaires dans les variables

Lorsque les paramètres de régression sont sous forme linéaire et les variables explicatives sous forme non-linéaire, le modèle est globalement linéaire. Par exemple :

Y= $a_0 + a_1 \cdot \ln(X) + a_2 X + a_3 X^2 + \varepsilon$ est une régression linéaire, en posant $x_1 = \ln(X)$ et $x_2 = X^2$,

mais pas Y= $a_0 + X^{a_1} + X^{a_2} + \varepsilon$.

En d'autres termes on peut avoir une fonction curvilinéaire de la variable indépendante dans l'équation et considérer que le modèle est globalement linéaire. Dit encore plus simplement, avec une régression linéaire, on peut aussi décrire des courbes, comme dans les graphiques ci-dessous :



En partant du graphique en haut à gauche, et en allant dans le sens des aiguilles d'une montre, les courbes de régression pourraient avoir une formule de type :

$Y = a + bx + cx^2$ (quadratique)

$Y = a + b \ln(x)$ (logarithmique)

$Y = ab^{1/x}$

$Y = a + bx + cx^2 + dx^3$

On utilise souvent le logarithme lorsqu'on a des variables très asymétriques qui risqueraient de donner des résidus qui ne suivent pas une loi normale. C'est typiquement le cas pour des variables telles que le revenu, le salaire, et la fortune, car dans la plupart des pays il y a une petite minorité qui a des conditions de vie immensément supérieures à la moyenne.

On utilise souvent la forme quadratique lorsque le lien entre la variable explicative et la variable expliquée est en forme de U ou de U inversé. Par exemple, si un phénomène augmente entre 20 et 40 ans, puis diminue après, on introduira l'âge et l'âge au carré dans le modèle de régression.

10.4. Régressions multiples

Une régression multiple est de la forme :

$$y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_n x_n + \varepsilon,$$

c'est-à-dire que plusieurs variables expliquent la variable dépendante. Ce type de modèle est le plus couramment utilisé en sciences sociales ; en effet, il est rare qu'on puisse expliquer un phénomène de façon satisfaisante par une seule variable.

Ce type de modèle rajoute toutefois **une quatrième hypothèse technique : il ne doit pas y avoir de problème de multicolinéarite**, c'est-à-dire que les variables explicatives ne doivent pas trop être corrélées entre elles. Les logiciels informatiques calculent une grandeur appelée VIF (pour variance inflation factor) qui mesure la force du lien entre les variables explicatives. Si VIF est ≥ 5, il y a un problème important de multicolinéarité qui peut biaiser certains résultats.

11. RECAPITULATION : TESTS ET MESURES D'ASSOCIATION EN FONCTION DU TYPE DE VARIABLES ETUDIEES

	Nominale	**Ordinale**	**Quantitative**
Nominale	Test du χ^2/ coefficient de contingence, coefficient de Cramer	Tests de Kruskal-Wallis ou de Mann-Whitney (**la variable expliquée est ordinale**)	ANOVA/ R^2 (**la variable expliquée est quantitative**)
Ordinale		ρ de Spearman et τ de Kendall / Tests : ρ ou $\tau = 0$	ANOVA/ R^2 (**la variable expliquée est quantitative**)
Quantitative			Régression simple/ Coefficient de Pearson (**la variable expliquée est quantitative**)