CHAPITRE II

INTRODUCTION A L'ACTIVITE STATISTIQUE

Semaine épouvantable : pas un seul sondage d'opinion.
Tant pis, nous essaierons de deviner tout seuls nos propres intentions.
André Frossard – Les pensées

1 - LA DEMARCHE STATISTIQUE

Sans vouloir donner de points de repères historiques précis (voir pour cela l'annexe correspondante), on peut affirmer que l'activité statistique est vieille comme le monde, ou plutôt comme l'Etat. En effet, l'étymologie nous indique que le terme même vient du latin moderne *statisticus* (relatif à l'Etat) formé à partir du latin classique *status* (Etat). Pour toutes les langues latines, le mot désignant la statistique a été formé sur cette racine. Mais les anglo-saxons n'étaient pas en reste puisque jusqu'au XVIII^{e} siècle, le terme consacré était celui de *political arithmetic* qui tomba ensuite en désuétude au profit de *statistics*. Bien sûr, les choses ont considérablement évolué depuis puisque la statistique a largement dépassé ce cadre et n'épargne que très peu d'activités humaines (sciences et technologie, commerce et économie, sports, politique…), mais la terminologie est restée marquée par ces origines comme le montrent les lignes qui suivent.

Aujourd'hui le terme statistique recouvre plusieurs sens qu'il s'agit de bien différencier. Tout d'abord, la statistique correspond au sens originel de l'activité, c'est-à-dire que la statistique est constituée d'un ensemble des méthodes qui ont pour objet la collecte, l'analyse et l'interprétation de données relatives à un ensemble d'unités de même nature. Les unités en question sont désignées par le terme d'**individus** et l'ensemble de ces unités constitue la **population**.

Le terme de statistique désigne également l'objet de l'activité précédente, à savoir l'ensemble des informations recueillies sur la population concernée. Il s'agit du sens courant du terme, abondamment employé dans les media.

Enfin, lorsque des observations sont faites sur la population entière ou seulement sur un **échantillon**, la statistique désigne toute fonction des observations calculée de manière à donner un résumé de l'information obtenue. Nous verrons plus en détail comment interpréter précisément le terme d'échantillon, mais on peut déjà dire à ce stade qu'un échantillon est une sous-population de la population étudiée formée de façon la plus représentative possible. Mais toutes ces acceptions vont se clarifier à travers les deux exemples qui suivent, le premier proche des préoccupations originelles de la statistique et le second dans un contexte industriel.

Exemple 1 – sondage électoral

Paul Hitique est candidat à la mairie de Saint Etienne. A un mois des élections, il désire savoir si sa position sur les questions de sécurité est bien perçue par les électeurs. Pour ce faire, il s'adresse à une société spécialisée en sondages d'opinion, la société LC2, et lui demande de bien vouloir quantifier la proportion de personnes qui sont d'accord avec sa position. La société LC2 fait de la statistique au premier sens du terme, en mettant les moyens nécessaires (humains, financiers, méthodologiques) pour répondre à la question posée. Ce faisant, les **individus** concernés sont les électeurs de la ville de Saint Etienne et la **population** celle de ces électeurs. On notera par exemple que les étrangers ou les mineurs, non électeurs, ne font pas partie de la population. Pour effectuer son étude, la société LC2 ne pourra évidemment pas interroger l'ensemble des individus et devra se contenter de recueillir l'opinion d'une fraction de la population, à savoir un échantillon. Elle complétera d'ailleurs la question posée par M. Hitique en examinant également l'influence de diverses variables (sexe, age, catégorie socio-professionnelle, quartier…) sur la réponse fournie[1]. L'ensemble des données recueillies constitue une statistique au second sens du terme. La société LC2 fournira ainsi des tableaux et graphiques du type suivant :

age / sexe	18 à 30	31 à 40	41 à 50	51 à 60	61 à 70	plus de 71
femme	38%	42%	45%	45%	53%	61%
homme	42%	44%	48%	50%	55%	60%

pourcentage d'individus en accord avec la position de M. Hitique

dont une expression graphique peut être :

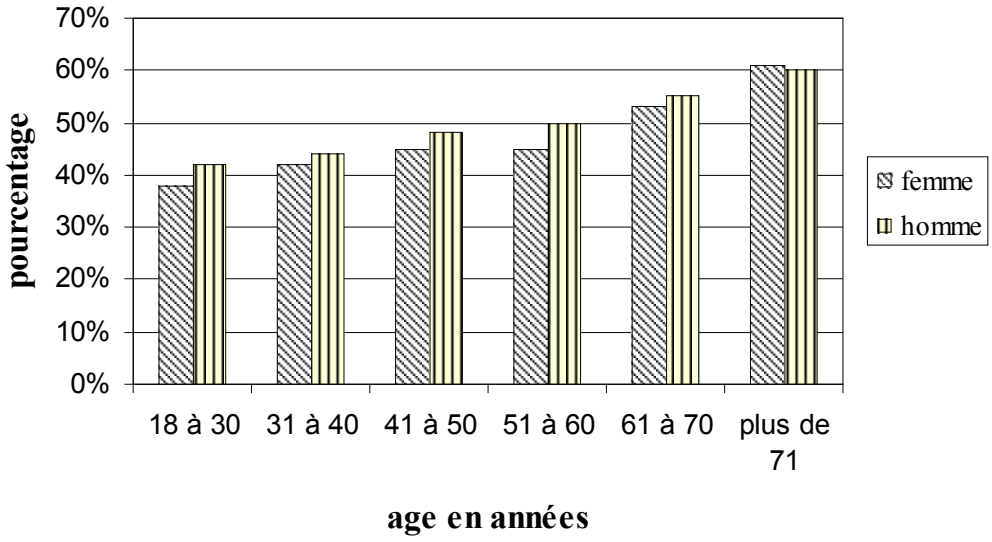

Figure 1. pourcentage d'individus en accord avec la position de M. Hitique

[1] La première tâche du statisticien consiste toujours à préciser, deviner, voire suggérer (!), les questions que se posent son interlocuteur.

Nous n'aborderons pas encore le problème de représentativité des chiffres fournis par l'étude. On comprendra très simplement que ceux-ci n'ont pas la même pertinence s'ils ont été obtenus à partir d'un échantillon constitué de 50 individus, de 1000 individus, de la famille du PDG de la société LC2… ou de celle de M. Hitique !

Pour terminer avec ce problème électoral, une statistique au troisième sens de la définition est par exemple le pourcentage d'individus de l'échantillon en accord avec la position de M. Hitique. Pour voir cela, modélisons un peu.

Supposons que les personnes interrogées soient numérotées de 1 à n, et désignons par ε_i la réponse de la personne numéro i : si celle-ci pense comme M. Hitique, on pose $\varepsilon_i = 1$. Dans le cas contraire, on pose $\varepsilon_i = 0$. A l'aide de ces notations, le pourcentage p de personnes de l'échantillon en accord avec la position de M. Hitique n'est autre que la moyenne :

$$p = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i$$

La quantité p, en tant que fonction des observations, est donc une statistique au sens troisième de notre définition.

Venons-en maintenant à un exemple plus industriel.

Exemple 2 – contrôle de production

M. Saint Quesse est nouvellement nommé à une fonction de responsable qualité dans une entreprise fabriquant des engrenages de transmission de puissance. Ayant effectué un premier tour d'horizon de l'entreprise, il s'aperçoit que le contrôle effectué sur la qualité des objets produits est plutôt rudimentaire et désire suivre de façon plus précise la qualité des productions au cours du processus de transformation. Dans cette optique, il cherche par exemple à caractériser le pas circulaire[2] d'un engrenage cylindrique droit.

Pour ce faire, il désire connaître les variations de ce pas au cours de la fabrication afin de repérer des dérives dans la production par exemple[3]. Les **individus** auxquels il s'adresse sont ici les engrenages produits et la **population** est celle de tous les engrenages produits et *à produire*. Le désir de connaissance de M. Saint Quesse se concrétisera par une démarche statistique au premier sens du terme puisqu'il s'agira pour lui d'utiliser des méthodes de collecte, d'analyse et d'interprétation des pas circulaires mesurés sur les individus engrenages. Dans ce contexte, on comprendra aisément qu'il ne peut être question d'interroger tous les individus de la population puisque les engrenages à venir ne sont pas encore produits ! En d'autres termes, il faudra se contenter de mesurer les pas circulaires d'un certain nombre d'engrenages, constituant un **échantillon**. Evidemment, il faudra pour cela s'assurer que l'échantillon choisi est représentatif de la production. Nous y viendrons plus tard. Dans l'immédiat, en supposant avoir recueilli les données correspondant à un échantillon de 150 engrenages, nous cherchons à reproduire l'information obtenue. Tout d'abord, un extrait du tableau des valeurs est reproduit ci-dessous.

[2] Il s'agit de l'espace entre deux dents mesuré sur le cercle constituant l'engrenage.

[3] Pour des compléments, voir le paragraphe consacré au contrôle de qualité situé au chapitre 7.

numéro	pas circulaire en cm
1	4,498
2	4,521
3	4,534
4	4,538
5	4,531
6	4,514
7	4,523
8	4,535
9	4,537
10	4,529
11	4,537
12	4,527
13	4,515
14	4,509
15	4,526
16	4,525
17	4,508
18	4,534
19	4,533
20	4,518

pas circulaire mesuré : 20 premières valeurs obtenues

Ce faisant nous effectuons de la statistique au second sens du terme en reproduisant ces données. Néanmoins, il est difficile de caractériser un tel tableau de chiffres sans analyse supplémentaire. De la même manière, le graphique le plus simple que l'on puisse faire, à savoir un tracé des valeurs obtenues en fonction du numéro de la pièce analysée est assez peu explicite.

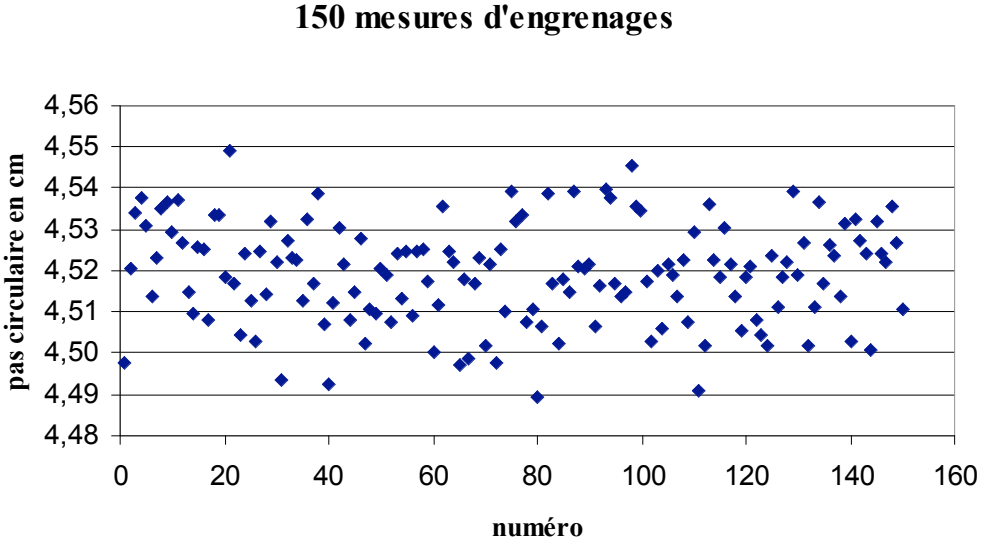

Figure 2. mesures du pas circulaire des 150 engrenages

La production de tableaux et de graphiques plus lisibles, et surtout plus informatifs[4], sera l'objet de la partie statistique exploratoire à venir. Pour en terminer

[4] qui seront obtenus en perdant de l'information car… trop d'information tue l'information !

temporairement sur cet exemple, des statistiques au troisième sens de la définition sont constituées de quantités variées. Par exemple, en notant p_i le pas obtenu pour l'engrenage numéro i ($1 \leq i \leq 150$), on peut exhiber les quantités suivantes, qui cherchent à caractériser le comportement global de l'échantillon étudié :

la moyenne $\bar{p} = \frac{1}{n} \sum_{i=1}^{n} p_i = 4{,}519$

le pas maximal $p_M = \max \{p_i, i = 1,\ldots, 150\} = 4{,}549$

le pas minimal $p_m = \min \{p_i, i = 1,\ldots, 150\} = 4{,}489$

Les étapes d'une étude statistique

Une première analyse des deux exemples qui précèdent nous montre que l'on peut grossièrement distinguer les étapes suivantes :

1. Collecte des données.
2. Statistique exploratoire.
3. Inférence statistique.

La collecte des données s'attache à définir, puis recueillir, les données pertinentes pour l'analyse des questions posées. La phase qui suit vise elle à explorer ces données afin de mieux comprendre la réalité qu'elles recouvrent. La phase inférentielle cherche enfin à fournir des outils rationnels d'aide à la décision, permettant d'indiquer que tel facteur dans une expérience est influent, que tel groupe de personnes a une opinion vraiment différente de tel autre, etc… Evidemment, cette classification est très rustique et la pratique fait que ces phases, en particulier les deux dernières, sont très souvent mêlées car la compréhension de jeux de données complexes passe par de nombreuses décisions qui sont fournies par la statistique inférentielle. Mais commençons par découvrir en détail les deux premières de ces phases, la dernière faisant l'objet de tous les autres chapitres de nature statistique présents dans cet ouvrage !

2 - COLLECTE DES DONNEES

Cette phase représente, et de loin, la plus grosse part d'une étude statistique car elle peut nécessiter la mobilisation de nombreuses personnes, mais aussi éventuellement de matériels expérimentaux complexes. Nous l'avons découpée en 4 parties.

2.1 - Définition de l'objet d'étude

La nécessité d'une telle définition est largement évidente car, avant de collecter des données, il faut savoir de quoi l'on parle. Aussi étonnant que cela paraisse, ce premier travail est souvent traité trop rapidement alors que son effet sur les résultats est de grande importance. Pour se convaincre de la nécessité et de la difficulté d'une telle tâche, examinons rapidement l'exemple qui suit.

Vous êtes employé dans le service des eaux d'une grande ville, et plus particulièrement affecté à la gestion du réseau d'évacuation des eaux usées. Le réseau est

construit de telle manière que lorsque de très fortes pluies s'abattent sur la ville, les différents collecteurs débordent dans la rivière traversant la ville, provoquant de nombreuses pollutions. C'est pourquoi vous êtes chargé de recueillir des statistiques sur les pluies afin à court terme de prévoir les épisodes difficiles, et à moyen terme de redimensionner le réseau. On vous demande donc concrètement de fournir des statistiques sur les pluies.

La question paraît très simple et parfaitement définie car rien n'est plus palpable que la pluie qui tombe, d'autant que la ville est équipée de quelques pluviomètres. Pourtant, un peu de réflexion montre que le problème ne se pose pas aussi simplement pour de nombreuses raisons. Par exemple, on imagine que la saison doit être un facteur important, de même que l'heure de la journée car le réseau d'eaux usées n'est certainement pas dans le même état tout au long de la journée, ce qui signifie que deux pluies identiques peuvent avoir des effets différents. Mais la question la plus délicate est sans doute celle qui paraît la plus simple, à savoir la définition d'une pluie ! En effet, les pluviomètres vont donner des courbes ressemblant à celle qui suit.

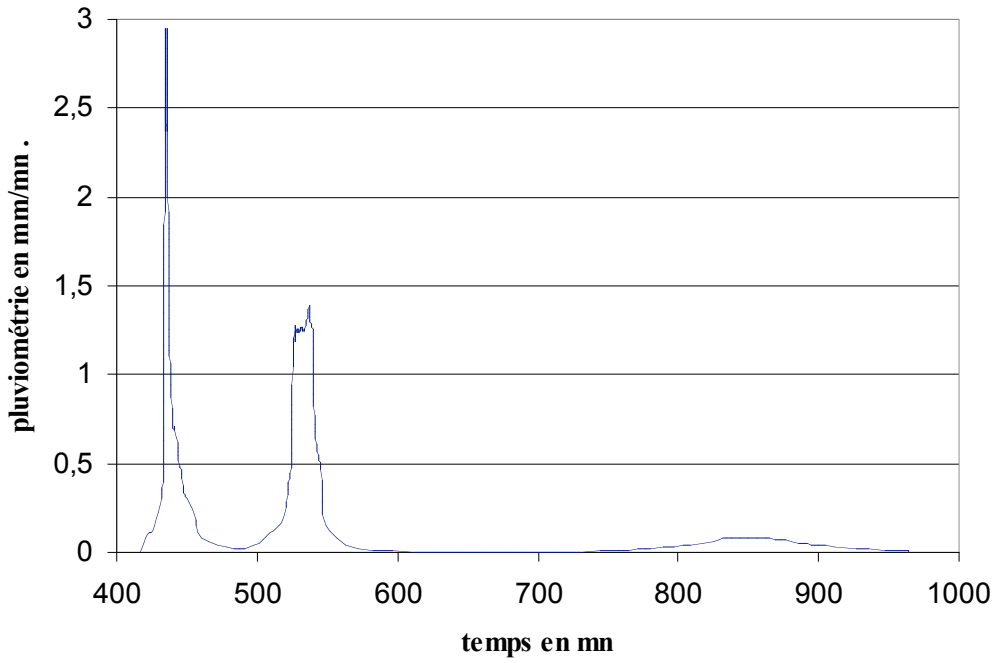

Figure 3. Pluviométrie relevée au cours d'une journée

Remarquons tout d'abord que les temps antérieurs à la 400ème minute et postérieurs à la 1000ème ne sont pas reproduits du fait qu'aucune précipitation n'a été relevée au cours de ces périodes. De nombreuses questions se posent alors. Par exemple, combien de pluies ont été observées au cours de la journée ? Quelle ont été leurs durées respectives ? leur intensité ? Attachons-nous ici à donner quelques éléments de réflexion pour la première question. Du fait que l'écart entre les deux premiers pics est faible, il n'est pas évident a priori de décider si l'on a observé 2 ou 3 pluies au cours de la journée. Il n'existe pas de réponse générale à cette question car tout dépend en fait de l'utilisation des données. Ici, le

facteur important est le temps nécessaire pour qu'un épisode pluvieux soit évacué par le réseau, appelé *temps caractéristique*. Si ce temps est très faible, de l'ordre de 30 mn, on aura intérêt à séparer les deux pics en considérant que l'on a eu deux épisodes pluvieux, l'un entre les temps 417 et 461, le second entre les temps 497 et 572. Si par contre le *temps caractéristique* du réseau dépasse par exemple 90 mn, on devra considérer la présence d'un unique épisode pluvieux entre les temps 417 et 572 car le réseau n'aura pas eu le temps d'évacuer les eaux de la première pluie au moment où la seconde arrive. A l'extrême, en négligeant les phénomènes orageux, un agriculteur considère que le nombre et la qualité de phénomènes pluvieux est sans importance pour ses cultures et que seule la pluviométrie cumulée sur une journée est à considérer puisqu'elle est une bonne estimation de la quantité d'eau rendue disponible pour les plantes. On voit par là que la phase de définition de l'objet d'étude est en fait une première activité de modélisation, au sens où nous l'avons définie au chapitre précédent, puisque celle-ci est intimement liée à l'application visée et aux questions posées.

Remarquons pour terminer que cette phase de définition des *individus* est évidemment liée étroitement à celle de la *population* d'intérêt, ici par exemple celle des pluies tombées au cours d'une certaine période passée, **ainsi que** les pluies à venir pour un horizon à préciser.

2.2 – Recensement ou sondage

Les individus et la population étant bien définis, il s'agit maintenant de préciser si l'on va recueillir les informations pour tous les individus de la population, on parle alors de **recensement**, ou seulement d'une partie de celle-ci auquel cas nous allons pratiquer un **sondage**. Il est très rare que l'on puisse effectuer un recensement, et ceci pour des raisons variées.

Il existe tout d'abord de nombreuses configurations dans lesquelles la population considérée n'est pas observable entièrement. C'est le cas de l'exemple qui précède où la population contient des individus *pluie* qui seront observés seulement dans les mois ou les années à venir.

Même dans le cas où la totalité de la population est observable, de nombreuses considérations de coût, financier, humain, organisationnel, mènent à effectuer un sondage. C'était le cas de notre exemple introductif relatif au sondage électoral. Et ç'aurait pu être également celui de l'exemple en contrôle de production si l'on avait considéré que la population était restreinte aux engrenages déjà produits[5]. On notera sur ce thème que comme dans toute activité humaine, une notion de rapport coût/efficacité est souvent en filigrane des choix effectués. Car un sondage peut être plus précis qu'un recensement ! Reprenons l'exemple de sondage électoral pour s'en convaincre. Toute personne ayant réalisé des enquêtes sait qu'une bonne gestion des réponses aux questions posées est souvent très délicate, et que certaines questions peuvent être mal interprétées. Pour réaliser une enquête de bonne qualité, une formation des enquêteurs est nécessaire, et de nombreux débriefings doivent être organisés afin que, au minimum, les processus de recueil des données soient homogènes entre les enquêteurs. On comprend bien dans ces circonstances qu'un sondage réalisé sur 1000 personnes par 10 enquêteurs bien formés,

[5] Notons au passage le cas extrême du contrôle destructif de produits où un recensement est évidemment absurde !

qui font des points réguliers, va donner des résultats de meilleure qualité qu'une enquête nécessitant 200 enquêteurs devant interroger 200 000 individus. Et nous n'évoquons pas le fait que dans une telle situation, certains des 200 enquêteurs auront trouvé un autre travail en cours d'enquête… De plus, nous verrons par la suite que dans le cas d'un échantillon bien construit, on peut facilement quantifier la précision du résultat, alors que les errements que nous venons d'évoquer sont évidemment difficilement traduisibles en erreur quantifiée sur le résultat.

2.3 – Echantillonnage

Cette partie ne concerne évidemment que le cas, de loin le plus courant, où il a été décidé de procéder par sondage dans la population. Il s'agit alors de préciser la ou les méthodes à employer pour effectuer la sélection de l'échantillon. Et la question n'est pas aussi simple qu'il pourrait le sembler comme le montre le célèbre exemple de l'élection présidentielle aux USA de 1936 opposant MM. Roosevelt et Landon. Les sondages d'opinion étaient développés dans ce pays depuis le début du XIX° siècle. Il s'agissait de *votes de paille*[6] réalisés sans véritable méthodologie (bulletin à découper dans des journaux, urnes installées à la sortie de bureaux, journalistes interrogeant des passants dans la rue…). Afin de prévoir le résultat de cette élection présidentielle, la revue *Literary Digest* réalise un vote de paille à partir de listes d'abonnés au téléphone et de propriétaires de voitures, obtenant plus de 2 millions de réponses. Les résultats obtenus donnent une confortable victoire pour Landon avec environ 54% des votes. Par ailleurs, G.H. Gallup, récent fondateur de l'*American Institute of Public Opinion*, prévoit – sur la base d'un échantillon de quelques milliers de personnes – une victoire pour Roosevelt avec un score d'environ 56%. Roosevelt gagnera finalement l'élection avec 62,5% des voix, et la fortune de G.H. Gallup s'en suivra ! La différence essentielle entre les deux approches est que Gallup a constitué son échantillon de manière la plus représentative possible de la population américaine alors que la consultation effectuée par le Literary Digest sur-représente les populations aisées.

Afin de constituer un échantillon représentatif de la population, on distingue aujourd'hui deux grandes classes de méthodes, les méthodes probabilistes et les méthodes empiriques. Les premières ont l'avantage de fournir des informations sur la précision des estimations obtenues alors que les secondes ont, comme leur nom l'indique, des bases moins solides mais sont néanmoins souvent utilisées pour des raisons de coût ou d'organisation.

Randomisation

Il s'agit de la méthode de sondage probabiliste conceptuellement la plus simple, à partir de laquelle les autres méthodes probabilistes sont construites. Si la population est constituée d'individus notés $\{i_1, i_2,\ldots, i_N\}$, on va chercher à tirer un échantillon de taille n ($1 \leq n \leq N$) de manière uniforme parmi tous les échantillons possibles. On peut montrer qu'un tel tirage peut être effectué en tirant uniformément le premier individu dans la population, puis uniformément le second parmi la population restante, et ainsi de suite

[6] L'expression provient de l'image du paysan jetant une paille dans le vent pour en connaître la direction.

jusqu'au n-ième individu[7]. On parle dans ce cas de tirage sans remise dans la population, et on montre que chaque individu a la même probabilité d'être choisi.

Le principe d'une telle méthode est extrêmement simple et applicable dans des circonstances très variées. On peut cependant se heurter à des problèmes de mise en œuvre, et l'exemple 2 de contrôle de production en fournit une illustration. Si l'on désire réaliser un échantillon randomisé de taille 100 de la production d'une journée, il est peu raisonnable de demander à un opérateur d'attendre pour observer successivement des pièces dont les numéros ont été tirés au hasard, par exemple les numéros 7, 19, 21, 45, 56, 85…

Par ailleurs, afin d'améliorer la précision des estimations obtenues, il est souvent intéressant de modifier la randomisation. La méthode qui suit en donne un exemple.

Stratification

Plaçons-nous dans une situation où l'on cherche à estimer la proportion de personnes travaillant dans une entreprise venant au travail sans utiliser leur véhicule automobile (transports en commun, cycles, rollers, voyage à pied…). Il est bien évident que la distance entre le domicile et le lieu de travail va influencer grandement le mode de transport utilisé. Si l'on sait par exemple que 40% des personnes habitent à plus d'un km de leur lieu de travail et que l'on désire interroger 10 personnes, on va naturellement choisir au hasard 4 personnes habitant à plus d'un km et 6 à moins d'un km. On comprend bien qu'un tel choix donnera des résultats moins variables qu'une randomisation sur l'ensemble de la population. En effet, par randomisation, on pourrait au contraire obtenir un échantillon de 6 personnes travaillant à plus d'un km et 4 à moins d'un km[8]. Dans ces conditions, l'estimation obtenue pour la proportion de personnes n'utilisant pas leur automobile sera sensiblement diminuée et représentera mal la population.

Pour formaliser, nous dirons qu'une **stratification** est basée sur la définition de groupes d'individus, appelés strates, notés S_1, …, S_k. Ces groupes forment une partition de la population, et chaque strate S_i représente une proportion p_i de la population. Dans ces conditions, un échantillon stratifié de taille n[9] est constitué en effectuant un sondage par randomisation d'environ $n\,p_i$ individus dans chaque strate S_i.

Il est intuitivement évident que la qualité de cette forme de sondage dépend fortement de la justesse des paramètres p_i. Si ceux-ci sont mal connus, l'estimation faite à partir du sondage stratifié sera entachée de biais. Nous verrons au chapitre consacré aux qualités des estimateurs comment formaliser et quantifier ce raisonnement intuitif.

Méthode des quotas

Cette technique de sondage, que nous citons ici du fait de sa popularité, ressemble à la stratification sans être pour autant une méthode probabiliste. Il ne passe en effet pas un jour sans que l'on lise dans un quotidien une phrase du type : *sondage réalisé auprès d'un échantillon de 1012 individus selon la méthode des quotas*. Les quotas en question sont les proportions de chaque strate du sondage stratifié. En pratique pour les sondages d'opinion, les strates les plus utilisées sont les fameuses catégories socio-professionnelles

[7] Voir plus loin la loi hypergéométrique.

[8] La probabilité d'un tel événement est d'environ 0,11.

[9] On parle en fait de sondage stratifié à allocation proportionnelle.

(CSP), remplacées d'ailleurs depuis 1982 par l'INSEE par les professions et catégories socioprofessionnelles, dites PCS. La seule différence entre la méthode des quotas et la stratification provient de la méthode de sondage dans chaque strate. Dans le cas de la méthode des quotas, le tirage aléatoire dans chaque strate est remplacé par des consignes données à l'enquêteur pour respecter une certaine représentativité des individus interrogés. Cette méthode permet en pratique de réduire le coût de l'enquête au prix d'une qualité de l'estimation moins bonne, et surtout non quantifiable.

Sondage systématique

Cette méthode consiste à représenter la population sous la forme d'une liste, par exemple du type $\{i_1, \ldots, i_n\}$, puis à construire l'échantillon en deux temps. On tire le premier individu au hasard dans la liste - notons-le i_k -, puis on tire les numéros des individus suivants selon un pas p déterminé à l'avance. Le second individu est donc l'individu i_{k+p}, le troisième i_{k+2p}, etc… Evidemment, lorsque l'on dépasse le numéro maximal n, on retourne au début, travaillant en fait en arithmétique modulo n. Cette méthode est très utilisée en contrôle de production, où p prend la signification d'un intervalle de temps, pour des raisons logistiques à peu près évidentes. Sa qualité s'approche de celle de la randomisation à condition que la qualité des individus ne soit pas dépendante du temps. Pour bien comprendre l'écueil potentiel, il suffit d'imaginer une personne estimant le trafic routier entre Lyon et Saint Etienne en se plaçant tous les jours à la même heure ! Le choix du pas d'échantillonnage doit donc être effectué avec soin.

2.4 - Dépouillement, traitement des données aberrantes.

Nous ne ferons ici que signaler l'existence de cette phase, fondamentale en pratique, qui représente souvent un travail conséquent. Car les données recueillies ne sont jamais parfaites, toutes dans les mêmes unités, sans erreur de saisie et le travail de mise en forme et contrôle conditionne grandement l'étude ultérieure. Nous passons également ici sous silence la question délicate des données manquantes qui devient cruciale lorsque la dimension des données traitées augmente. Toutes ces questions sont souvent réalisées en parallèle avec la statistique exploratoire qui suit.

3 - STATISTIQUE EXPLORATOIRE

Pour reprendre le vocabulaire général, les observations faites sont basées sur des individus, chacun d'eux étant décrit par des variables. Celles-ci peuvent être de natures différentes, et on comprend bien que leur traitement ultérieur devra en tenir compte. Une variable peut en effet être de l'un des types suivants :

- **qualitative nominale**
  Dans ce cas, la variable peut prendre plusieurs modalités non comparables entre elles. Par exemple, si l'individu est un actif, une variable nominale le décrivant peut être sa catégorie socio-professionnelle (CSP)
- **qualitative ordinale**

Ici, la variable peut également prendre plusieurs modalités, qui peuvent être ordonnées. Par exemple, si l'on s'intéresse aux stages effectués par les étudiants ingénieurs au cours de leur scolarité, divers avis de l'entreprise sur le déroulement du stage peuvent être recueillis. Dans ce cas, l'individu est un étudiant et, concernant notamment l'autonomie de l'étudiant au cours de son stage, l'entreprise pourra choisir entre les modalités « très autonome », « a besoin d'un appui dans les phases clés de son travail », « sollicite un peu trop souvent de l'aide pour progresser », « dépendant ». Dans un tel cas, la variable « avis », tout en étant qualitative, peut être ordonnée. On notera que la question d'un codage numérique éventuel est alors particulièrement délicate car le passage d'un niveau codé 1 à un niveau codé 2 ne correspond pas au doublement d'une caractéristique de l'individu.

- **quantitative discrète**
  La variable considérée est dans ce cas caractérisée par un nombre, mais ce nombre ne peut prendre que des valeurs discrètes, assez souvent en nombre fini. On peut par exemple considérer le cas où l'individu est une famille et la variable le nombre d'enfants de celle-ci.
- **quantitative continue**
  Ici, la variable peut prendre un continuum de valeurs. C'est le cas que nous avons rencontré au cours de l'exemple 2 : l'individu était un engrenage produit et la variable le pas circulaire de ce dernier.

Comme nous avons pu commencer à le démontrer à propos de l'exemple 2, les données brutes recueillies sont dans la quasi-totalité des cas inexploitables en l'état. Il est en effet particulièrement difficile d'analyser des tableaux de chiffres de grande dimension sans chercher à résumer l'information sous-jacente. D'où la nécessité pour la personne chargée de ce travail d'utiliser des méthodes variées, reposant toutes sur des hypothèses. En effet, nous verrons dans ce qui suit que faire un graphique ou donner la moyenne d'une série de nombres n'est pas anodin et correspond à une activité où la subjectivité intervient toujours !

Nous structurerons notre propos entre résumés graphiques et résumés numériques. On notera que nous concentrerons notre propos sur des résumés concernant une variable à la fois. L'obtention de nombres ou de graphiques concernant des données multidimensionnelles sortent du cadre de cet ouvrage, concentré sur les questions mono-dimensionnelles. Nous renvoyons par exemple à Saporta (1990) pour un exposé sur ces questions, en ce qui concerne la statistique exploratoire.

3.1 - Résumés graphiques

Ils dépendent évidemment du type de variable considérée.

Variables qualitatives

Nous sommes ici dans le domaine de prédilection d'un logiciel comme Excel et renvoyons à ce dernier pour les multiples possibilités de représentation offertes : « histogramme[10] », camembert, anneaux…

Variables discrètes

Les valeurs observées sur l'échantillon sont notées x_1, ..., x_n et les valeurs possibles a_1, ..., a_k. Les effectifs observés pour chaque valeur possible sont :

$$n_j = \text{Card}\{i, x_i = a_j\}$$

et les fréquences observées :

$$f_j = n_j/n$$

➢ DIAGRAMME EN BATONS[11]

C'est le diagramme des couples (a_j, n_j) ou (a_j, f_j), les valeurs a_j étant placées sur l'axe réel. Nous prendrons comme exemple des données provenant d'une enquête réalisée dans le département du Nord pour l'Observatoire Départemental de la Famille[12]. Divers renseignements ont été recueillis auprès de 224 élèves de classes de quatrième issus de 5 collèges différents. Le nombre d'enfants vivant au domicile familial est une des variables observées, le diagramme en bâtons correspondant est reproduit ci-après.

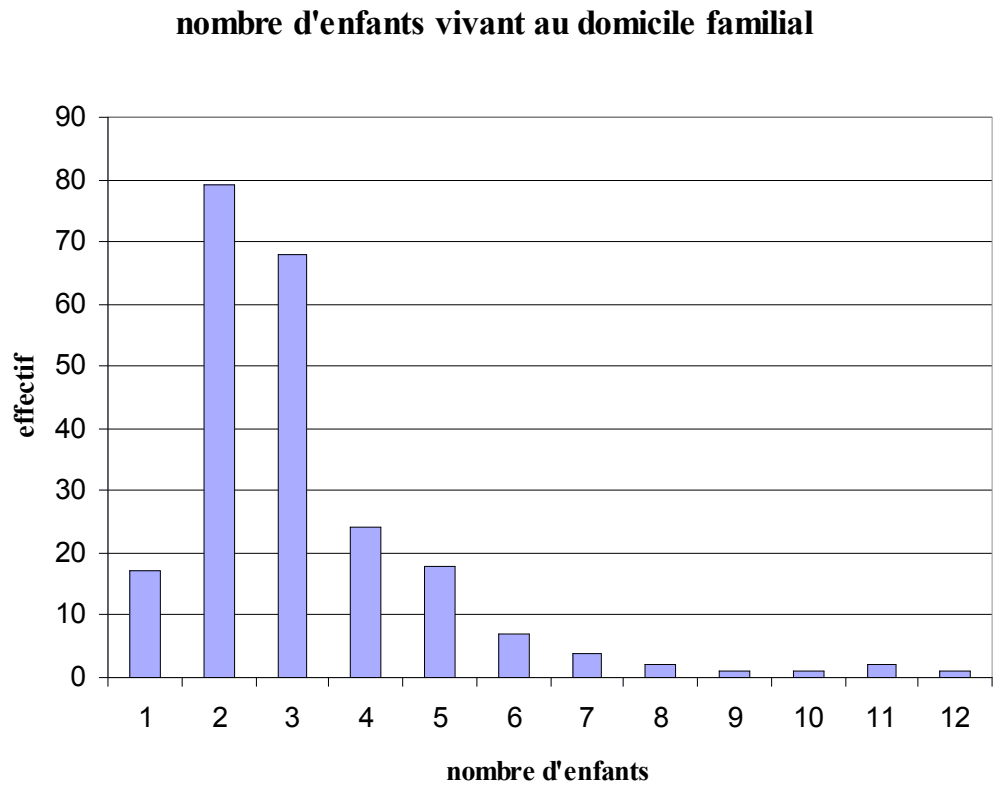

Figure 4. exemple de diagramme en bâtons

[10] Il s'agit en fait de diagrammes en bâtons, voir ci-après.

[11] Le diagramme en bâtons est également valable pour des variables qualitatives.

[12] L'enquête est intitulée « Quels sont les liens existant entre le contexte familial et l'absentéisme scolaire de l'adolescent ? ».

Comme dans tout travail de représentation, il convient de bien réfléchir à ce que l'on désire observer avant de choisir un mode de représentation. Par exemple, si l'on désire comparer pour cet exemple deux populations de collégiens de tailles différentes, on choisira préférentiellement de représenter les fréquences observées plutôt que les effectifs.

➢ FONCTION DE REPARTITION EMPIRIQUE

C'est une fonction correspondant aux effectifs cumulés, ou plus précisément aux fréquences cumulées.

La **fonction de répartition empirique** de l'échantillon $(x_1, \ldots, x_n)$ est la fonction F_n définie par :

$$\text{pour } x \in \mathbb{R},\ F_n(x) = \frac{1}{n}\,\text{Card}\{i,\ x_i \leq x\}$$

Dans l'exemple qui précède, le graphe de cette fonction donne la figure suivante.

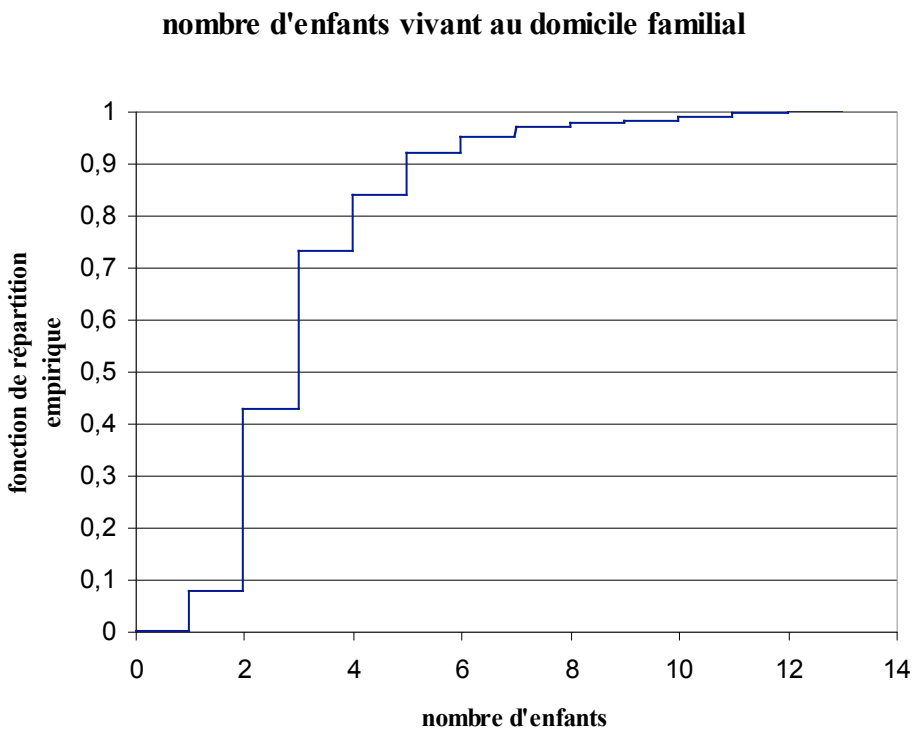

Figure 5. exemple de fonction de répartition empirique

Une telle représentation offre des possibilités différentes de celles du diagramme en bâtons. Il est par exemple très facile ici de lire directement sur le graphique le pourcentage d'élèves pour lesquels le nombre d'enfants vivant au domicile familial est inférieur ou égal à 4 ; on obtient environ 84%.

Variables continues

On suppose ici avoir observé les valeurs $x_1, \ldots, x_n$.

➢ FONCTION DE REPARTITION EMPIRIQUE

La définition est identique à celle du cas discret. Dans le cas de l'exemple de contrôle de production et de mesure de pas d'engrenages, le graphique est le suivant.

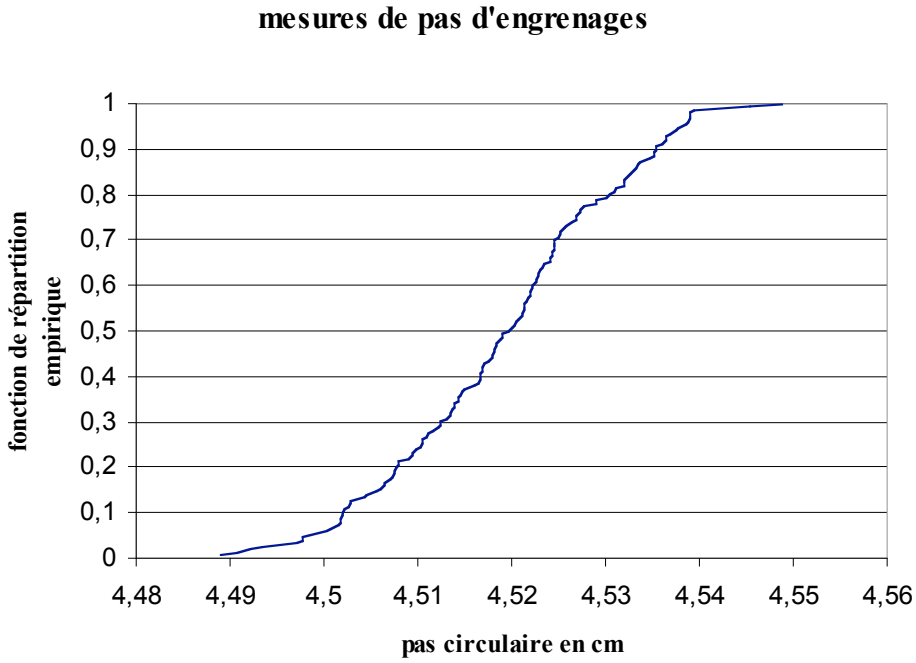

Figure 6.

➢ HISTOGRAMME

Contrairement au cas de la fonction de répartition empirique dont la définition n'est pas affectée par le caractère discret ou continu de la variable observée, il n'est pas facile de construire pour une variable continue un graphique analogue au diagramme en bâtons. C'est d'ailleurs la raison pour laquelle nous avons commencé par définir la fonction de répartition empirique. Pour pouvoir construire un tel graphique, qui aura pour fonction de représenter une certaine *densité* de données en chaque point, l'idée la plus simple consiste à regrouper les observations en classes. On notera ainsi[13] :

$$C_1 = [a_0, a_1[, ..., C_k = [a_{k-1}\ a_k[$$

Par analogie au diagramme en bâtons, on définit la fréquence d'appartenance à la classe C_j :

$$f_j = \frac{1}{n} \mathrm{Card}\{i, x_i \in C_j\}$$

Mais l'analogie s'arrête là car il s'agit ici de représenter une densité de données, c'est-à-dire une fréquence de données par unité de *longueur*. C'est pourquoi on représente alors une fonction en escaliers dont la valeur sur la classe C_j est $f_j/(a_j - a_{j-1})$. Cette représentation a l'avantage de fournir une aire sous la courbe égale à 1 puisque l'aire sous la marche d'escalier associée à la classe C_j est égale à la fréquence f_j. De plus, elle rend bien compte lorsque les largeurs de classes sont variables du fait que, à fréquence constante, une classe deux fois plus petite qu'une autre se verra attribuée une densité deux

[13] On suppose évidemment que toutes les données sont comprises entre a_0 et a_k.

fois plus forte. Dans la pratique, les classes sont très souvent de largeur constante h et un bon choix de h est délicat comme le montre l'exemple qui suit.

Dans le cas des pas d'engrenages, un premier histogramme construit sur des classes régulières, de largeur 0,005, débutant en $a_0 = 4{,}485$ donne la figure suivante.

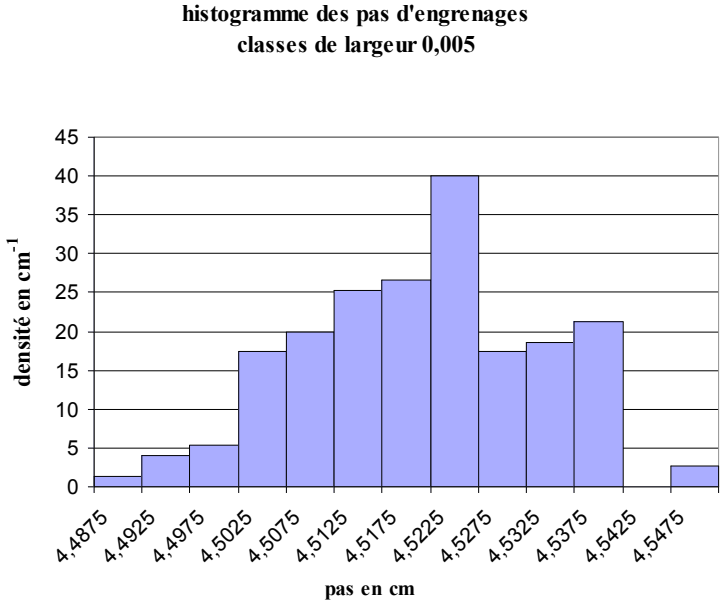

Figure 7.

On notera au passage que nous avons ici représenté chaque classe par son milieu. Ce choix a été effectué du fait que Excel[14] ne peut pas en l'état construire un histogramme avec une variable réelle en abscisse. Ce qui rend d'ailleurs en l'état impossible la construction sous Excel d'histogrammes formés à partir de classes de tailles différentes. Abordons maintenant l'effet de la largeur h sur la représentation obtenue. Un examen des deux cas extrêmes qui suivent montre les bornes entre lesquelles doit se situer une représentation graphique satisfaisante.

[14] Nous ne saurons trop avertir le lecteur sur les nombreux pièges que contient Excel. Par exemple, une traduction erronée de l'anglais *frequency* en fréquence laisse à penser que Excel calcule les fréquences observées dans chaque classe alors qu'il s'agit des effectifs.

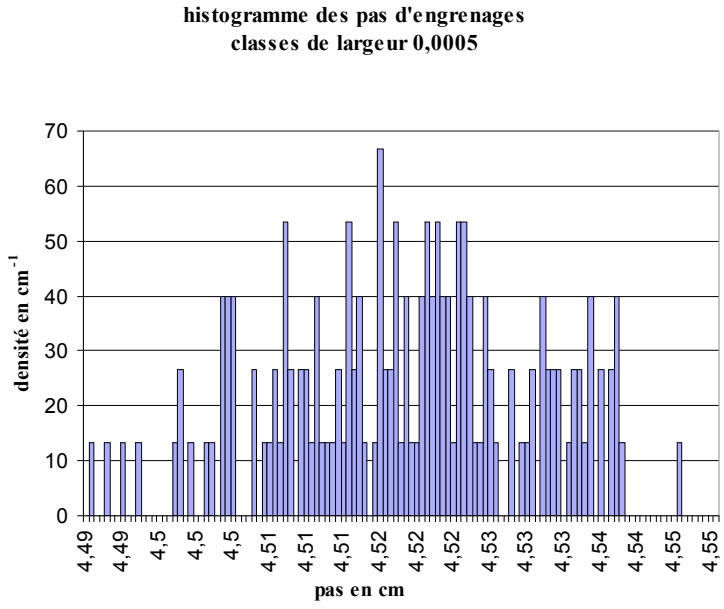

Figure 8.

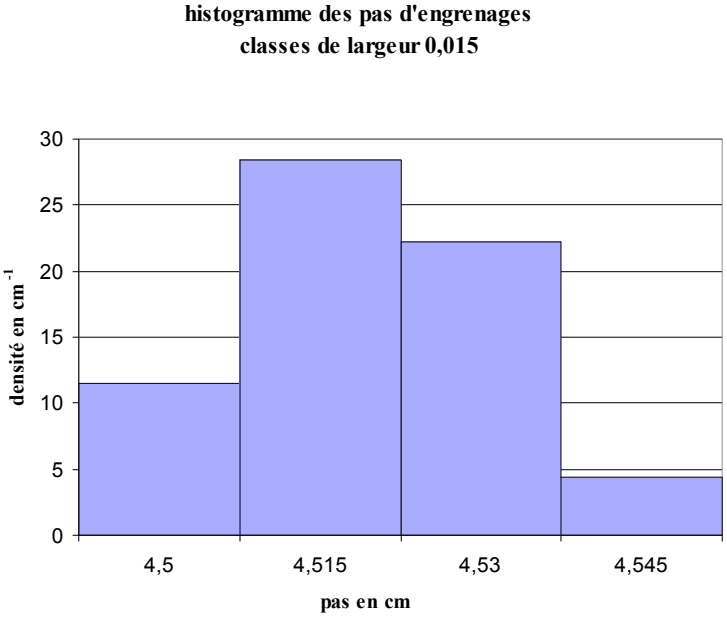

Figure 9.

Le premier graphique correspondant à des classes de largeur faible est très difficilement interprétable, car très "bruité" du fait que très peu de données sont présentes dans chaque classe. Par contre, la largeur des classes étant faible, on perd très peu d'informations sur les données en réalisant ce graphique. A contrario, le second graphique correspond à des classes très larges, donc peu nombreuses. Le graphique obtenu est donc très peu perturbé mais fait perdre énormément d'informations. L'objectif d'une bonne représentation est donc de perdre le moins d'informations possibles, tout en conservant un graphique lisible.

Autant les considérations précédentes peuvent être étendues à presque toute représentation graphique[15], autant l'histogramme possède des inconvénients qui lui sont propres. Le problème le plus important de l'histogramme est sa non robustesse au choix de la largeur de classe. Pour s'en convaincre, il suffit de comparer la figure ci-après – correspondant à une largeur de 0,004 – à la première reproduite qui correspond elle à une largeur de classe de 0,005. Une faible variation du paramètre de largeur donne des graphiques sensiblement différents. Nous aurions également pu effectuer la même remarque à propos du choix de la borne inférieure de la première classe, choisie par l'opérateur, qui souffre du même problème de sensibilité.

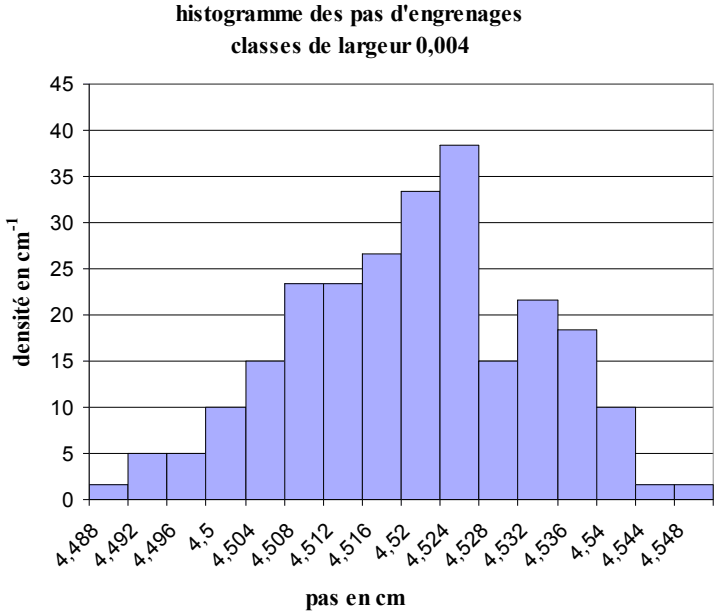

Figure 10.

➢ METHODES DE LISSAGE

Un peu de réflexion permet de se convaincre que les problèmes de non robustesse de l'histogramme à certains paramètres proviennent du fait que l'on modifie le choix des classes, une donnée peut brusquement changer de classe d'appartenance. L'écriture qui suit va à la fois nous en convaincre et nous donner des voies d'amélioration. En effet, si l'histogramme est construit sur des classes C_1, …, C_k de largeur constante égale à h, l'estimation en un point x appartenant à la classe C_j de l'histogramme est donnée par :

$$\hat{f}(x) = \frac{1}{nh}\,\mathrm{Card}\{i,\ x_i \in C_j\} = \frac{1}{nh}\sum_{i=1}^{n} K(x, x_i),$$

où $K(x, x_i)$ vaut 1 si x_i appartient à même classe que x et 0 sinon.

Le problème rencontré avec cette estimation est lié à la discontinuité de la fonction K, qui provoque des sauts brusques dans l'expression lorsque h, et donc les classes C_j, varie. Une façon de résoudre ce problème consiste simplement à utiliser une fonction K,

[15] On notera cependant que la fonction de répartition empirique ne fait perdre aucune information.

appelée noyau, qui soit continue. Ce qui revient en fait à donner un poids à chaque donnée qui varie continûment, au lieu de ne prendre que les valeurs 0 ou 1 pour l'histogramme.

Nous profiterons de cette liberté pour introduire des expressions explicitement invariantes par translation, c'est-à-dire non modifiées lorsque le point courant x et les données $(x_i)_i$ sont translatés d'un même facteur. Tout cela nous mène à la définition qui suit.

LISSAGE PAR NOYAU

Si K est une densité de probabilité[16], le lissage des données par le noyau K, de largeur de fenêtre h, est la fonction $\hat{f}$ définie par :

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$$

Remarquons que de même que pour l'histogramme, la fonction $\hat{f}$ est construite de manière à ce que son intégrale soit égale à un[17]. Par suite, la probabilité (ici la proportion) d'appartenance à un intervalle [a,b] peut être estimée par $\int_a^b \hat{f}(x)\,dx$.

Revenons au données précédentes pour examiner les graphiques obtenus. Nous utiliserons pour cet exemple un noyau K gaussien, i.e. la densité de la loi normale N(0,1)[18]. Remarquons tout d'abord que le choix de la largeur de fenêtre est aussi délicat que dans le cas de l'histogramme, avec les mêmes caractéristiques, comme le montrent les deux graphiques qui suivent.

[16] C'est-à-dire une fonction positive d'intégrale 1, voir le chapitre suivant pour des détails sur cette importante notion.

[17] Exercice immédiat.

[18] Pour une expression de cette densité, voire les tables en annexe.

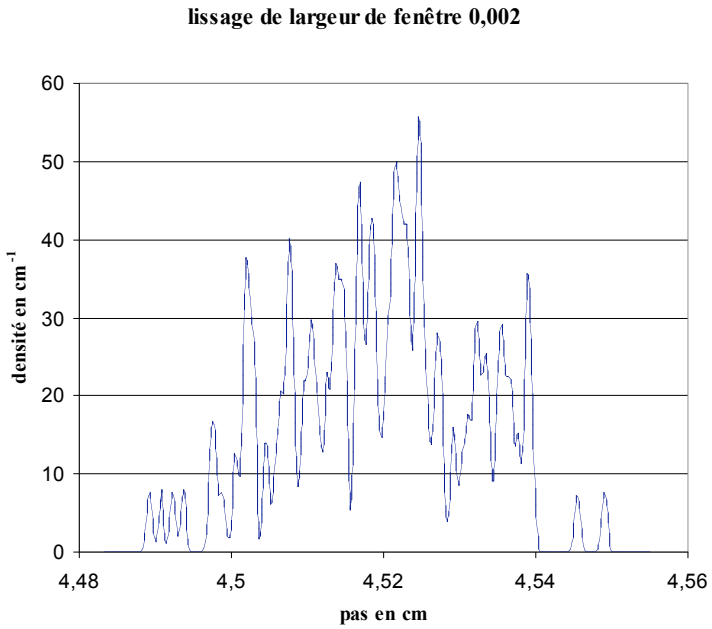

Figure 11.

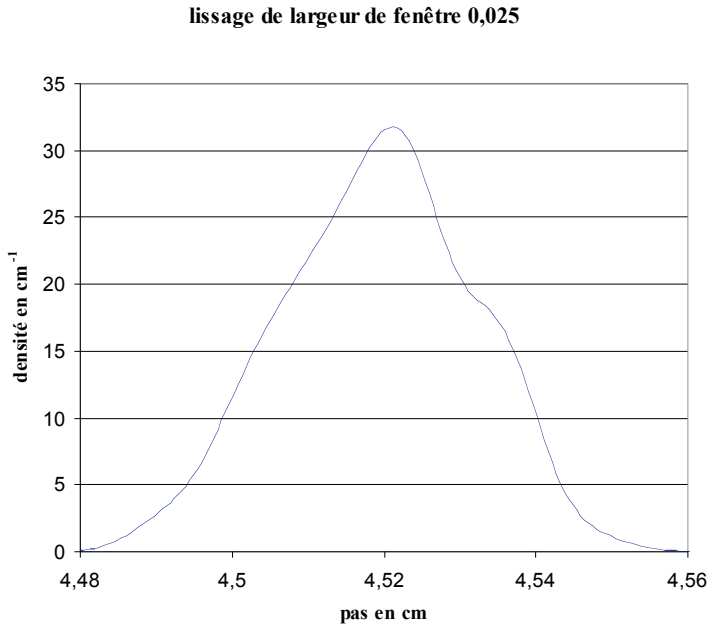

Figure 12.

Les phénomènes déjà observés pour l'histogramme sont à nouveau présents ici même si la courbe obtenue est toujours de classe C^∞. Entre ces deux extrêmes se situent les largeurs de fenêtre assurant à la fois une bonne représentation des données et une fonction dont les variations restent modérées. C'est ce qu'illustrent les deux graphiques qui suivent.

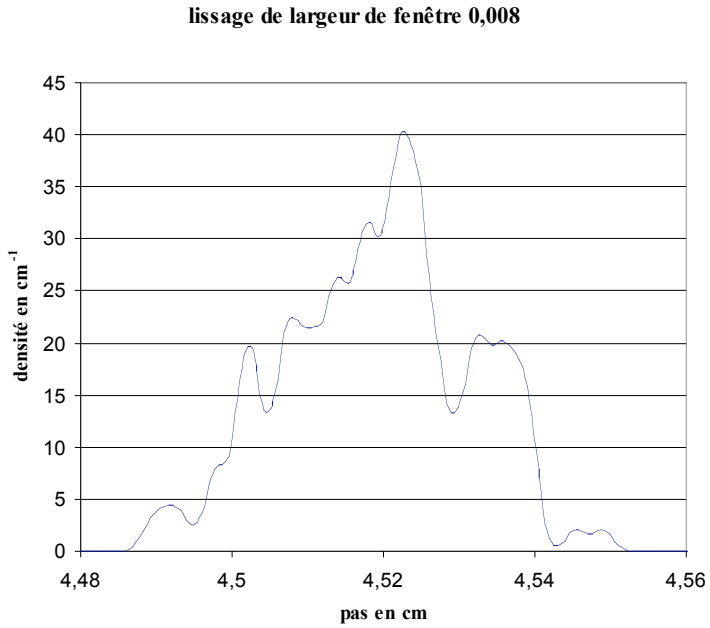

Figure 13.

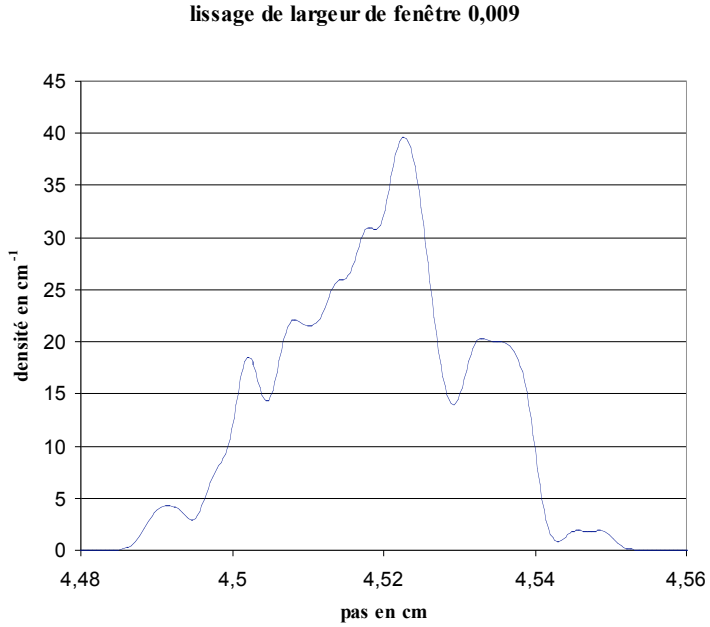

Figure 14.

On notera ici que, contrairement au cas de l'histogramme, une faible variation de la largeur de fenêtre donne une faible variation du graphique obtenu. Ce qui constitue un avantage essentiel de cette méthode.

En conclusion, nous ne pouvons que conseiller au lecteur d'éviter d'utiliser l'histogramme dont l'intérêt est surtout historique[19] et de représenter plutôt la fonction de répartition empirique ou le lissage par noyau.

[19] Contrairement au lissage par noyau, l'histogramme pour de petits échantillons peut être construit « à la main ».

3.2 - Résumés numériques

Nous ne considérerons ici que le cas des variables quantitatives, discrètes ou continues. Les données recueillies constituent un échantillon noté $x_1, \ldots, x_n$.

CARACTERISTIQUES DE TENDANCE CENTRALE

L'objectif poursuivi ici est d'exhiber un seul nombre qui représente le mieux possible la valeur autour de laquelle les données sont situées. La première idée est évidemment celle de la moyenne.

MOYENNE

La moyenne de l'échantillon $x = (x_i)_i$, notée $\bar{x}$ est définie par :

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

La propriété la plus importante, et pourtant la plus évidente, est la suivante.

LINEARITE DE LA MOYENNE

Si $x = (x_i)_i$ et $y=(y_i)_i$ sont deux échantillons de taille n et a, b deux réels, on a :

$$\overline{ax+by} = a\,\bar{x} + b\,\bar{y}$$

Plus généralement, on notera que la plupart des calculs algébriques sont facilités par cette linéarité. Supposons par exemple que l'on acquiert les données au fil du temps et que la donnée x_i est acquise au temps $i\,\Delta t$. Notons alors $\mathbf{x}^n$ la série de données acquise entre les temps 0 et $n\,\Delta t$, c'est-à-dire $\mathbf{x}^n = (x_1, \ldots, x_n)$. Un calcul très simple montre que la mise à jour de la moyenne lorsque l'on passe du temps $n\,\Delta t$ au temps $(n+1)\,\Delta t$ est obtenue par combinaison barycentrique de la moyenne précédente $\bar{\mathbf{x}}^n$ et de la nouvelle donnée acquise x_{n+1} :

$$\bar{\mathbf{x}}^{n+1} = (1-\frac{1}{n+1})\,\bar{\mathbf{x}}^n + \frac{1}{n+1}\,x_{n+1}$$

On notera cependant que la moyenne est un indicateur peu **robuste**, c'est-à-dire sensible aux valeurs extrêmes. Un simple exemple numérique permet de s'en convaincre. Reprenons les données des 150 pas circulaires d'engrenages et évaluons la moyenne. On obtient :

$$\overline{pas} \approx 4{,}519$$

Supposons maintenant que les données soient saisies par un opérateur et que celui-ci fasse une erreur, par exemple en modifiant la position de la virgule. On peut imaginer pour fixer les idées qu'il remplace la première donnée 4,498 par 44,98. L'effet sur la moyenne de cette erreur sur l'une des 150 données est important puisque l'on obtient alors :

$$\overline{pas} \approx 4{,}789$$

On voit donc qu'une simple erreur de saisie pour une donnée sur 150 modifie la moyenne obtenue d'environ 6%.

Nous venons de voir que la moyenne n'était pas robuste, ce qui peut poser certains problèmes. Mais il est également possible que cet indicateur ne soit pas le plus adapté dans le contexte donné. Une alternative est fournie par la médiane.

MEDIANE

La médiane med(x) de la série de données $x = (x_i)_i$ est la valeur qui partage les données en deux parties égales. En d'autres termes, si on note $x_{(1)}$, ..., $x_{(n)}$ les valeurs ordonnées des x_i,

$\text{med}(x) = \frac{x_{n/2} + x_{(n/2)+1}}{2}$ si n est pair et $x_{(n+1)/2}$ si n est impair

Comme suggéré ci-avant, la médiane peut être plus adaptée au problème posé comme le montre l'exemple qui suit. Vous êtes un ingénieur nouvellement diplômé et cherchez un premier emploi. Vous hésitez entre les entreprises A et B, et souhaitez savoir s'il est plus intéressant, en termes de salaires à un horizon de 5 ans, de travailler dans l'entreprise A que dans l'entreprise B. Vous parvenez à obtenir les salaires d'ingénieurs en premier emploi dans ces entreprises depuis 5 ans. Vos observations sont résumées dans le tableau qui suit.

	Salaires annuels observés en k€							
Entreprise A	34500	37200	52300	36400	33800	34600	38000	
Entreprise B	37100	36900	39000	38500	36200	39400	37600	38300

Les moyennes et médianes de ces séries de données sont les suivantes

	Moyenne	Médiane
Entreprise A	38114	36400
Entreprise B	37875	37950

On observe que la moyenne des salaires de l'entreprise A est supérieure à celle de l'entreprise B, alors que la conclusion est inversée pour les médianes. Cela est dû à nouveau à la non robustesse de la moyenne, et ici au salaire très élevé d'un ingénieur de l'entreprise A. Pour le jeune diplômé, il est sans doute plus intéressant de savoir que la moitié des ingénieurs de l'entreprise A gagnent moins de 36400 € et la moitié gagnent davantage. Il peut ainsi se situer dans la population et estimer le salaire auquel il pourrait prétendre après 5 ans. De plus, l'ingénieur de l'entreprise A gagnant 52300 k€ représente sans doute un cas particulier en termes de carrière dont il est difficile de tenir compte sans information supplémentaire. En conclusion de cet exemple, la médiane est plus adaptée ici pour donner un salaire *typique* de la population considérée que la moyenne.

Comme nous l'avons vu dans l'exemple précédent, la médiane est beaucoup plus robuste que la moyenne, puisqu'elle ne voit pas les valeurs extrêmes : dès qu'une valeur dépasse la médiane, on peut la multiplier par un facteur d'échelle quelconque[20] sans que la médiane soit modifiée. Par contre, il est plus difficile d'effectuer des calculs algébriques à l'aide de la médiane. Même la relation de linéarité de la moyenne n'est pas conservée.

[20] A condition qu'il dépasse 1 !

Ainsi, se demander pour deux séries de données $x = (x_i)_i$ et $y=(y_i)_i$ si $med(x) > med(y)$ ne revient pas à s'interroger sur le signe de $med(x-y)$.

Enfin, la médiane est conservée par une transformation strictement croissante contrairement à la moyenne, ce qui peut rendre quelques services lorsque des transformations, logarithmique ou en puissance, sont appliquées sur les données considérées[21].

Beaucoup d'autres indicateurs peuvent être exhibés, par exemple des moyennes pondérées, les moyennes géométrique ou harmonique, les « trimmed mean » qui ôtent un certain pourcentage de valeurs aux extrémités de l'échantillon. Nous n'irons pas plus loin sur la question mais terminerons par un indicateur utilisable pour des données qualitatives.

MODE

Pour des données discrètes[22], **le mode** est la valeur la plus fréquemment observée dans l'échantillon.
Si les données sont continues, **le mode** est l'abscisse du maximum d'une estimation de la densité (histogramme, lissage par noyau…) et **un mode** est l'abscisse d'un maximum local d'une estimation de la densité.

On notera que pour des données continues, le mode est assez mal défini puisque dépendant d'une estimation de la densité. On utilise néanmoins souvent cette notion lorsque plusieurs modes apparaissent, on parle alors de distribution **multimodale**, ce qui est souvent un indice de mélange de populations. C'est d'ailleurs le cas des données d'engrenages pour lesquelles les lissages obtenus montrent plusieurs modes. Une lecture des figures 13 et 14 montre deux modes bien marqués situés aux environs des valeurs 4,523 et 4,533. Une analyse plus approfondie, nécessitant souvent un retour sur les données, peut alors permettre de décider si l'on est en présence d'un mélange de deux populations différentes, les origines pouvant être variables : changement d'opérateur, horaires de production, variation de matière première…

CARACTERISTIQUES DE DISPERSION

Une fois indiqué autour de quelle valeur les données sont situées, la question qui vient naturellement à l'esprit ensuite, fondamentale dans notre contexte statistique, est celui de la dispersion des valeurs, ou de leur plage de variation. Nous utiliserons abondamment la notion déjà rencontrée qui suit.

ECHANTILLON ORDONNE

L'**échantillon ordonné** construit à partir de l'échantillon $(x_1, \ldots, x_n)$ est la donnée des n valeurs ordonnées de façon croissante.
On le note $x_{(1)}, \ldots, x_{(n)}$

[21] Ce qui est souvent le cas lorsque l'on traite de problèmes de régression ou de séries temporelles, lorsque l'on désire stabiliser la variance.

[22] Cette définition s'étend aux données qualitatives.

ETENDUE

L'**étendue** est la quantité $x_{(n)} - x_{(1)} = Max(x) – Min(x)$

L'étendue est par nature non robuste puisque dépend des seules valeurs extrêmes. Elle est néanmoins utilisée, notamment en contrôle statistique de procédés.

➢ QUANTILES

L'idée sous-jacente à la définition consiste à introduire des quantités qui généralisent la notion de médiane pour une probabilité α différente de ½, i.e. des nombres qui partagent les données en deux parties de façon à ce que la proportion des données inférieures à cette quantité soit égale à α. Comme la fonction de répartition empirique F_n de l'échantillon donne précisément ces proportions, il est tentant de poser :

Pour $\alpha \in [0,1]$, le **quantile** d'ordre α - noté q_α - est donné par l'équation :

$$F_n(q_\alpha) = \alpha$$

On notera que cette définition n'en est pas véritablement une car l'équation en question possède en général soit aucune, soit une infinité de solutions, comme le montre la figure qui suit !

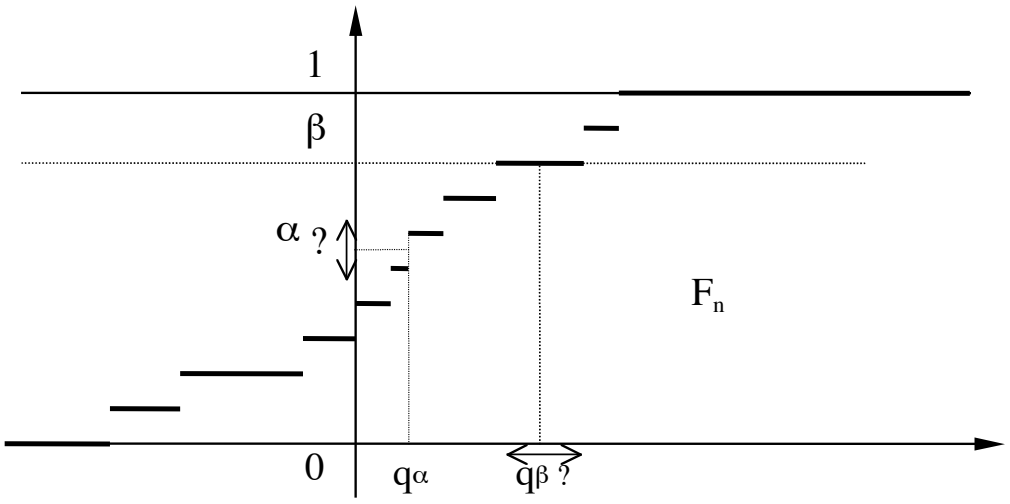

Figure 15.

Il va donc falloir choisir des conventions pour aller plus loin. Nous en choisirons une, parmi les plus courantes[23]. Elle consiste à définir les données extrêmes comme étant les quantiles d'ordres 0 et 1, puis répartir les données qui restent en n-1 intervalles d'égale probabilité, soit 1/(n-1). En utilisant l'échantillon ordonné $x_{(1)}, ..., x_{(n)}$, on obtient la :

[23] Cette convention est utilisée par Excel et R.

DEFINITION DES QUANTILES

Soit $\alpha \in [0,1]$, le quantile q_α est défini par[24] :

Pour i = 1, …, n-1 et $\alpha = \frac{i-1}{n-1}$, $q_\alpha = x_{(i)}$

Si α n'est pas de la forme précédente, q_α est défini par interpolation linéaire :

Si $\alpha = (1-\gamma)\frac{i-1}{n-1} + \gamma \square \frac{i}{n-1}$, avec $0 < \gamma < 1$, $q_\alpha = (1-\gamma)\, x_{(i)} + \gamma\, x_{(i+1)}$

Pour mieux comprendre cette définition, le plus simple est d'examiner le graphique ci-après, construit sur un exemple de faible dimension.

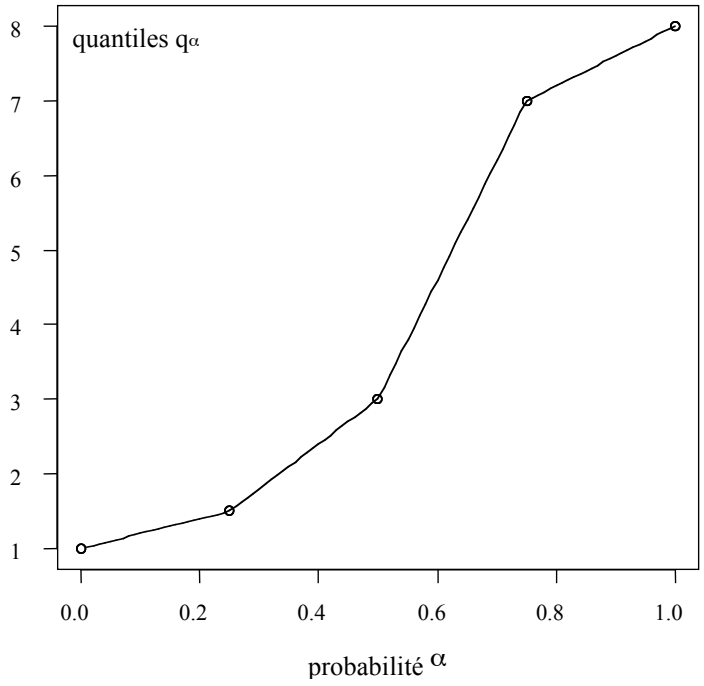

Figure 16. Quantiles de l'échantillon (1,1.5,3,7,8)

Le lecteur attentif pourra vérifier que le quantile d'ordre ½ n'est autre que la médiane. On utilise souvent des quantiles particuliers, par exemple les quartiles ($\alpha = i/4$, $1 \le i \le 3$), les déciles ($\alpha = i/10$, $1 \le i \le 9$), les percentiles ($\alpha = i/100$, $1 \le i \le 99$).

Les quantiles ont de nombreuses utilisations, en particulier pour nos questions de dispersion. Par exemple,

DISTANCE INTERQUARTILES

La distance interquartiles est la longueur de l'intervalle $[q_{1/4}, q_{3/4}]$, c'est-à-dire $q_{3/4} - q_{1/4}$.

Il est facile de se convaincre que cet indicateur est très robuste, mais compte tenu de sa définition, les manipulations algébriques sont difficiles, voire impossibles. On notera également que cette distance est facilement interprétable, puisque les 50% « centraux » des données appartiennent à un intervalle de cette longueur.

[24] La quantité (i-1)/(n-1) est remplacée dans la littérature par de nombreuses alternatives, par exemple i/n, i/(n+1), (i-1/2)/n… ou même (i-3/8)/(n+1/4).

Les quartiles sont également très utilisés pour réaliser des graphiques simples appelés *box-and-whisker plots*[25] ou boîtes à moustaches dont nous verrons la grande utilité pour comparer très simplement de multiples jeux de données. Ces graphiques sont construits en plusieurs étapes, dont le résultat est donné ci-après sur les données de pas d'engrenages :

1. Construction de la boîte. Elle est formée comme un rectangle d'extrémités les quartiles $q_{¼}$ et $q_{¾}$, coupée par un segment passant par la médiane med(x).
2. Construction des moustaches. On commence par définir, à l'aide de la distance interquartiles, les valeurs maximales que peuvent prendre les extrémités des moustaches. Il s'agit des valeurs min = $q_{¼}$ - 1,5 ($q_{¾}$ - $q_{¼}$) et max = $q_{¾}$ + 1,5 ($q_{¾}$ - $q_{¼}$). L'extrémité droite de la moustache est alors définie comme la plus grande donnée inférieure à max, et l'extrémité gauche comme la plus petite donnée supérieure à min.
3. Représentation des valeurs extrêmes. Il s'agit des données si elles existent qui ne sont pas situées entre les moustaches. Elles sont représentées par un symbole.

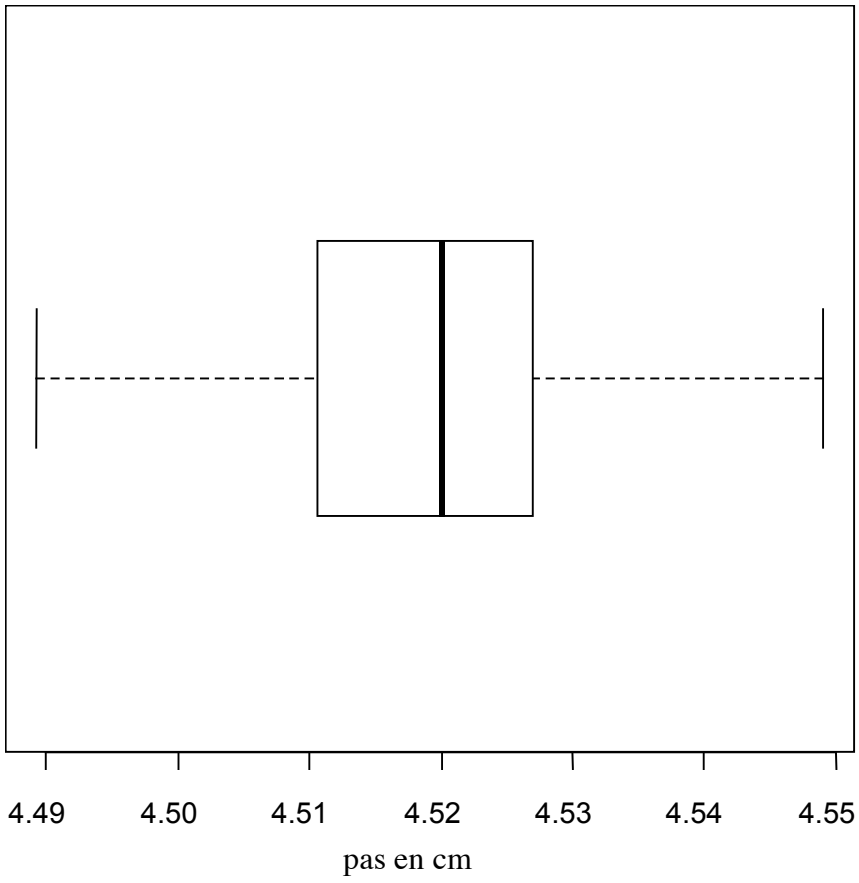

Figure 17. Box plot des 150 pas d'engrenages

On remarque sur la figure qui précède qu'aucune donnée n'est considérée comme extrême puisque aucun symbole n'est présent hors des moustaches, ce qui signifie que les extrémités des moustaches sont ici simplement le minimum et le maximum des pas observés. Nous verrons en fin de chapitre un exemple pour lequel apparaissent des valeurs

[25] On parle également de Tuckey plots en référence à leur inventeur.

extrêmes[26]. Mais avant cela, retrouvons les données de températures hebdomadaires du chapitre 1, regroupées par mois.

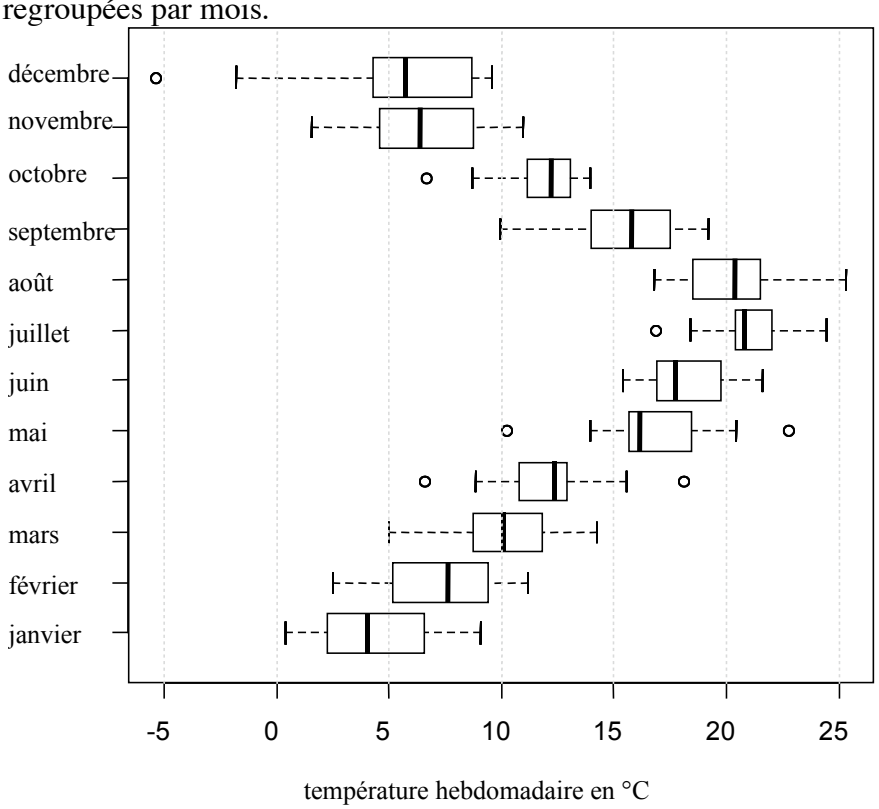

Figure 18. Box plot des températures hebdomadaires à Paris de 1996 à 1999

On observe ici une dépendance évidente de la température moyenne en fonction du mois, mais les différences entre les mois ne se limitent pas à un facteur de translation. Par exemple, on note également que la largeur des boîtes est plus importante en hiver qu'en été, ce qui signifie que la variabilité des températures dépend de la saison considérée. Nous n'irons pas plus loin sur l'interprétation[27]. Retenons simplement que les boîtes à moustaches sont un outils précieux pour la comparaison de populations.

Mais l'indicateur le plus courant reste à définir.

[26] Qualifier ces valeurs n'est pas une question simple et l'on trouve par exemple très souvent dans la littérature le qualificatif de valeurs aberrantes, laissant supposer que ces valeurs posent problème, ce qui est loin d'être toujours le cas.

[27] Nous conseillons cependant au lecteur de méditer l'adage « En avril ne te découvre pas d'un fil. En mai, fais ce qu'il te plait ».

VARIANCE, ECART-TYPE

Soit $x = (x_1, \ldots, x_n)$ un échantillon, la **variance** de x, notée Var(x) est définie par :

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

L'écart-type de x, noté σ_x, est donné par :

$$\sigma_x = \sqrt{\mathrm{Var}(x)}$$

L'idée de cet indicateur est de mesurer un *écart moyen* à la moyenne $\bar{x}$. Mais la moyenne des écarts à la moyenne est nulle, comme le montre un calcul immédiat. C'est pourquoi on cherche à comptabiliser les écarts positifs ou négatifs de la même manière, ce que permet la fonction quadratique. Mais du fait de l'apparition de termes carrés, la variance n'est pas exprimée dans la bonne unité. Par exemple, pour les pas d'engrenages mesurés en cm, la variance est donnée en cm^2. C'est la raison pour laquelle on introduit l'écart-type dont la valeur peut être comparée aux données. Cependant, on notera que, s'il est aisé d'interpréter l'étendue ou la distance interquartiles vis-à-vis des données, il n'en est pas de même de l'écart-type[28]. Il est clair également que l'écart-type n'est pas un indicateur robuste, car construit à partir d'une moyenne en élevant en outre les termes de la somme au carré, ce qui renforce la sensibilité aux valeurs extrêmes. Par contre, variance et écart-type jouissent de propriétés algébriques agréables. Par exemple, si x est un échantillon et a et b deux réels, on a :

$$\mathrm{Var}(ax+b) = a^2 \mathrm{Var}(x)$$

4 - INFERENCE STATISTIQUE

Comme nous l'avons vu en introduction de ce chapitre, la phase d'inférence a pour objet de fournir des outils d'aide à la décision qui visent tous peu ou prou à extrapoler à l'ensemble de la population des résultats observés sur l'échantillon. Pour ce faire, l'outil probabiliste va être prépondérant en assortissant l'ensemble de nos décisions ou indications de probabilités d'occurrence. Ce qui signifie évidemment que nos considérations statistiques vont faire appel à une modélisation probabiliste. La majeure partie des chapitres à venir va s'attacher à développer les outils ad hoc. Pour bien comprendre le type d'activité que nous allons développer, le plus simple est de poursuivre notre exemple de contrôle de production.

Supposons ici qu'après une étude préliminaire, résumée par les 150 données que nous avons traitées précédemment, nous observions un nouvel échantillon de la production. L'échantillon en question est de taille 39 et les valeurs obtenues sont reproduites dans le graphique qui suit, analogue à la figure 2.

[28] Nous verrons dans le chapitre suivant un lien entre écart-type et proportion de données dans un certain intervalle, avec l'inégalité de Bienaymé-Tchebychev.

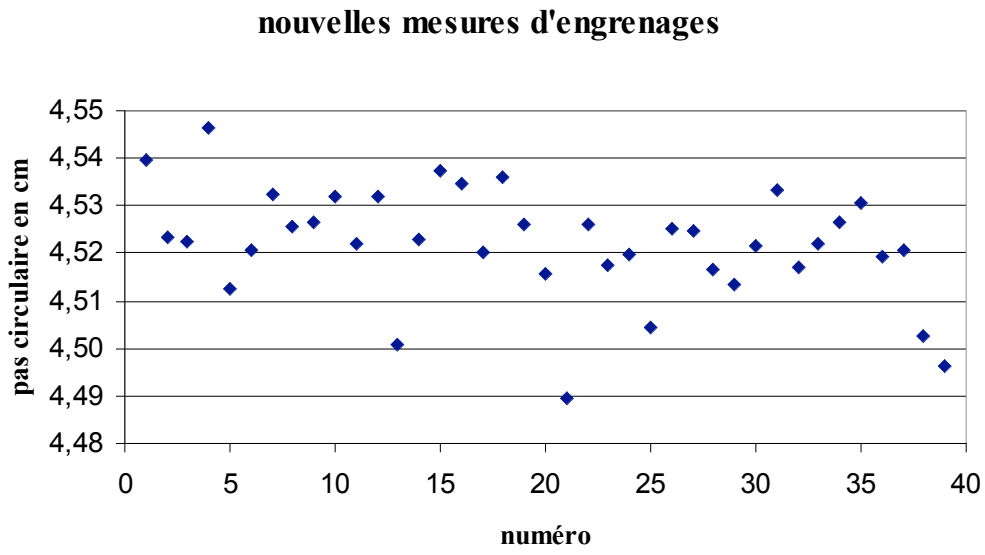

Figure 19.

Dans ce contexte, plusieurs questions se posent, dont nous retiendrons seulement les deux suivantes :

- Les nouveaux engrenages produits sont-ils différents des précédents ?
- Continuent-ils à respecter les tolérances imposées à la fabrication, en particulier que la moyenne des pas est comprise entre 4,519 et 4,521 ?

Pour comparer les deux échantillons, on peut comparer les estimations de densités par lissage :

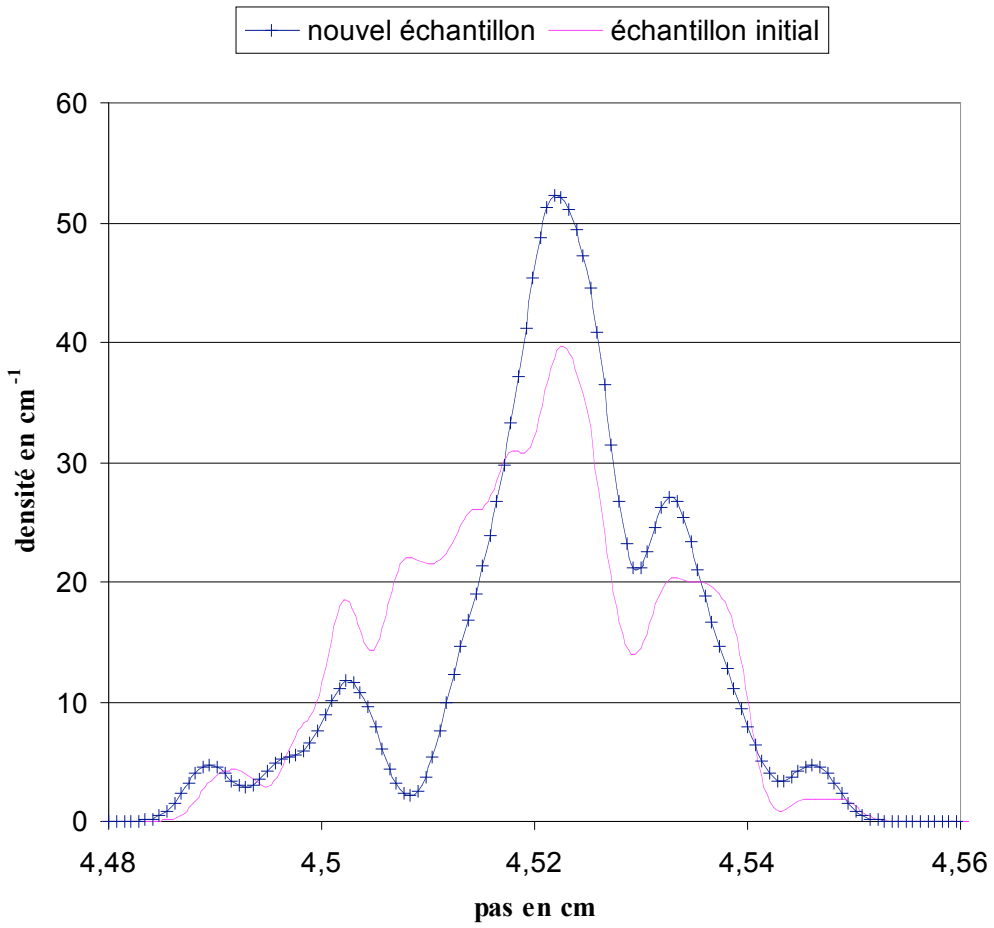

Figure 20. Comparaison des lissages des deux échantillons

Le résultat est un peu difficile à interpréter mais on observe néanmoins que le mode principal est estimé approximativement à 4,523 pour les deux échantillons tout en étant beaucoup plus marqué pour le second. De plus, la zone entourant 4,51 semble de probabilité beaucoup plus faible dans le cas du nouvel échantillon. Un autre type de comparaison est fourni par les boîtes à moustaches.

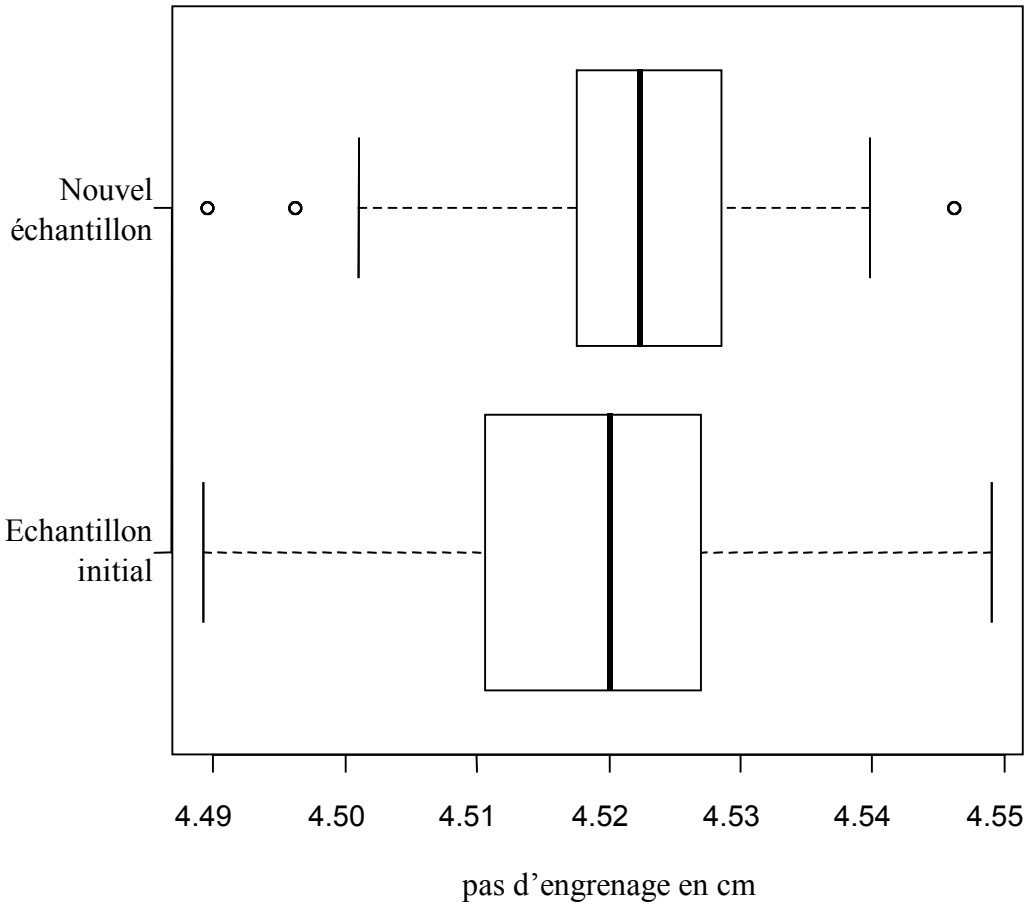

Figure 21. Boîtes à moustaches des deux échantillons

On admettra sans peine que cette nouvelle figure est beaucoup plus informative puisqu'elle montre plusieurs évolutions entre l'échantillon initial et le nouvel échantillon, à savoir un décalage de la boîte vers les valeurs fortes , une dispersion représentée par les moustaches moins importante, et la présence de quelques valeurs extrêmes. Ces remarques peuvent être consolidées par le tableau qui suit.

	moyenne	écart-type	médiane	premier quartile	troisième quartile	distance interquartile
échantillon initial	4,5192	0,0121	4,5201	4,5106	4,5270	0,0164
nouvel échantillon	4,5220	0,0116	4,5223	4,5174	4,5286	0,0112

On observe une augmentation sensible des indicateurs de tendance centrale, moyenne et médiane, ainsi qu'un décalage des quartiles. Les indicateurs de dispersion ont des comportements assez différents puisque l'écart-type varie de moins de 10% alors que la distance interquartile varie davantage.

Ces observations étant faites, il est difficile en l'état de répondre aux questions posées. On observe en particulier une moyenne dépassant la limite fixée de 4,52 et l'on pourrait s'inquiéter de la qualité de la production. Pourtant, il est important de noter que si la taille du nouvel échantillon était très faible (par exemple 10) ou très forte (de l'ordre de 1000), notre raisonnement ne serait pas modifié et notre inquiétude inchangée.

Pourtant, le bon sens nous indique que nos observations seraient plus assurées sur un échantillon de taille importante. Et un peu de réflexion nous convainc du fait que la moyenne observée sur un échantillon de 10 engrenages doit être très variable, donc d'interprétation délicate.

Aborder ces questions nécessitera tout d'abord de préciser le cadre probabiliste de l'étude. Il importe par exemple de préciser de quelle moyenne on parle lorsque l'on souhaite que celle-ci soit comprise entre 4,519 et 4,521. Cela nous mènera à la notion de variable aléatoire, de loi et d'espérance, questions abordées lors du chapitre 3. La notion d'intervalle de confiance nous permettra ensuite de dire si la moyenne observée sur l'échantillon correspond à une moyenne pour la population qui satisfait les contraintes de production. Ces notions, liées aux questions d'estimation de paramètres seront abordées au chapitre 4. Enfin, savoir si les moyennes de deux échantillons peuvent être considérées comme indistinguables nous conduira naturellement vers la théorie des tests au chapitre 7. Toutes les décisions prises seront assorties d'une probabilité qu'il s'agira d'interpréter avec soin.

5 - EXERCICES - ENONCES

○ II.1 – représentations graphiques

A partir du tableau de données ci-dessous, quel graphique vous paraît adapté

- pour représenter la superficie des régions françaises ?
- pour comparer la population estimée au 01/01/97 ?

Régions	**Superficie**	**Population estimée au 1.01.97**	**Produit intérieur brut (PIB) par habitant en 1996**	**Capitales régionales**	**Population (RP1990)**
Unités	**Km²**	**1 000 hab**	**KF**		**habitants**
Alsace	8280	1713	136	Strasbourg	252338
Aquitaine	41309	2891	120	Bordeaux	210336
Auvergne	26013	1315	109	Clermont-Ferrand	136181
Bourgogne	31582	1626	117	Dijon	146703
Bretagne	27209	2875	112	Rennes	197536
Centre	39151	2455	119	Orléans	105111
Champagne-Ardenne	25606	1352	122	Châlons-sur-Marne	48423
Corse	8680	262	106	Ajaccio	58949
Franche-Comté	16202	1118	120	Besançon	113828
Île-de-France	12011	11061	207	Paris	2152423
Languedoc-Roussillon	27376	2265	102	Montpellier	207996
Limousin	16942	719	105	Limoges	133464
Lorraine	23547	2311	115	Metz	119594
Midi-Pyrénées	45348	2520	113	Toulouse	358688

Nord-Pas-de-Calais	12414	4007	111	Lille	172842
Basse-Normandie	17589	1422	115	Caen	112846
Haute-Normandie	12318	1784	137	Rouen	102723
Pays de la Loire	32082	3173	118	Nantes	244995
Picardie	19399	1869	110	Amiens	131872
Poitou-Charentes	25809	1628	108	Poitiers	78894
Provence-Alpes-Côte d'Azur	31400	4478	119	Marseille	800550
Rhône-Alpes	43698	5646	130	Lyon	415487

II.2 – représentations graphiques

On s'intéresse à l'évolution de la population active en fonction du temps et on se base pour cela sur le tableau qui suit. Quel(s) graphique(s) construire si l'on veut comparer l'évolution des actifs dans chaque secteur d'activité ?

Année	**1970**	**1980**	**1990**	**1997**
Agriculture	2780	1843	1259	976
Industrie (sauf indust.manufac.)	2947	2787	2495	2203
Industrie manufacturière	4974	4849	4001	3451
Services marchands	6312	7658	9104	9627
Services non marchands	3887	4805	5619	6448
Total = population active	20900	21942	22478	22705

Répartition des actifs en France par secteurs (en milliers)

II.3 – taux d'inflation

Le tableau ci-dessous reproduit les taux d'inflation observés en France et au Japon entre 1970 et 2002. On cherche à résumer ces différents taux par un taux unique bien représentatif de la période. Quel taux choisiriez-vous parmi les suivants ?

taux moyen taux médian mode autre

Justifiez soigneusement votre réponse.

année	**taux d'inflation annuel en %**		**année**	**taux d'inflation annuel en %**	
	France	**Japon**		**France**	**Japon**
1970	5,9	7,6	**1987**	3,3	0,1
1971	5,5	6,4	**1988**	2,7	0,7

1972	6,2	4,8	**1989**	3,5	2,3
1973	7,3	11,7	**1990**	3,4	3,1
1974	13,7	23,1	**1991**	3,2	3,3
1975	11,8	11,7	**1992**	2,4	1,7
1976	9,6	9,4	**1993**	2,1	1,2
1977	9,4	8,2	**1994**	1,7	0,7
1978	9,1	4,2	**1995**	1,8	–0,1
1979	10,8	4	**1996**	2,1	0
1980	13,1	7,8	**1997**	1,3	1,7
1981	13,3	4,9	**1998**	0,7	0,6
1982	12	2,8	**1999**	0,6	–0,3
1983	9,5	1,9	**2000**	1,8	–0,9
1984	7,7	2,3	**2001**	1,8	–0,7
1985	5,8	2	**2002**	1,9	–0,9
1986	2,5	0,6			

II.4 – courses automobiles

Le tableau suivant donne les temps du meilleur tour pour les dix premiers du grand prix de Monaco 2001 (circuit de 3,378 km, 78 tours). On cherche à résumer ces différents temps par un temps unique bien représentatif. Quel temps choisiriez-vous parmi les suivants ?

temps moyen temps médian mode autre

tableau à ajouter

II.5 – fluctuations d'échantillonnage

On désire ici effectuer une étude empirique afin d'aborder la question cruciale des fluctuations affectant les quantités évaluées sur un échantillon de taille donnée. A cet effet, considérons le cas où l'on désire estimer, au sein d'une population de taille très grande, voire infinie, la proportion d'individus dotés d'une caractéristique donnée. Celle-ci peut être la réponse par oui ou par non à une question dans le cas d'un sondage d'opinion, mais aussi, dans le cas de notre exemple de contrôle de production, l'appartenance du pas d'engrenage à un intervalle fixé[29]. En définitive, il s'agit d'estimer la proportion p inconnue à partir d'un échantillon de n individus. Celle-ci sera estimée à partir de la proportion d'individus possédant la caractéristique observée dans l'échantillon. En d'autres termes, p est estimée par la quantité $\hat{p}$ définie par :

$$\hat{p} = \frac{\text{nb individus de caractéristique donnée}}{n}$$

[29] Il s'agit le plus souvent d'un intervalle dit de tolérance dans lequel la production peut fluctuer sans dégrader les caractéristiques fonctionnelles de la pièce produite.

Intuitivement, l'estimation $\hat{p}$ sera d'autant plus précise que la taille de l'échantillon est grande. Le but de l'exercice est d'examiner, à l'aide des outils de la statistique exploratoire, la manière dont les propriétés de $\hat{p}$ dépendent de la taille de l'échantillon. Pour ce faire, nous allons réaliser, un nombre n et une probabilité p étant fixés, un grand nombre - *n.simu* - d'échantillons de taille n, et évaluer la quantité $\hat{p}$ correspondante. Nos observations seront basées sur l'échantillon de taille *n.simu* formé des valeurs $\hat{p}$ observées et les investigations poussées aussi loin que possible.

6 - EXERCICES – INDICATIONS DE SOLUTIONS

Exercices II.1 et II.2

Il suffit de prendre le temps de réfléchir à ce que l'on désire représenter !

Exercice II.3

Comment calcule-t-on un taux d'inflation annuel sur une période de 2 ans ?

Exercice II.4

Se souvenir de la définition d'une vitesse !

Exercice II.5

On fixera une valeur pour *n.simu*. On pourra fixer ensuite quelques valeurs de p, par exemple une faible, une moyenne et une forte (0,1 – 0,5 – 0,9) en notant que du fait que les échecs pour p = 0,1 sont les succès pour p = 0,9, il suffit par exemple d'examiner les cas 0,1 et 0,5.
L'utilisation, pour chaque valeur de p, de boxplots pour examiner la dépendance des résultats en fonction de la taille de l'échantillon est profitable.
Une étude ultérieure de la dépendance de la moyenne et de l'écart-type en fonction de cette même taille peut être menée assez loin.