TD Mathématiques et statistiques_CM_2022-2023

Année universitaire 2022-2023

Planning des séances

Séance 1. Introduction et rappels de cours.
¤ Présentation du fonctionnement (règles et évaluation) et du déroulement du cours.
¤ Introduction - la socio et les stats.
¤ Définitions (population, variable, modalité ; effectifs) ; pourcentages, taux de variation et tableaux statistiques.

Séance 2. Les variables.
¤ Note de participation : QCM ou mots croisés sur le vocabulaire de la séance précédente
¤ Cours : les différents types de variables et les échelles de mesure.

Séance 3. Construire des tableaux statistiques et représentations graphiques.

Séance 4. Contrôle continu 1.

Séance 5. Paramètres de position.
¤ Paramètres de position 1 : le mode et la moyenne
¤ Correction du contrôle et remise des devoirs

Séance 6. Paramètres de position.
¤ Paramètres de position 2 : la médiane et les quartiles

Séance 7. Paramètres de dispersion.
¤ La variance et l'écart-type

Séance 8. Révisions.

Séance 9. Contrôle continu 2.

Séance 1. Introduction, vocabulaire et statistiques

Présentation du fonctionnement et du déroulement du cours

¤ Déroulement général du cours.

Cet enseignement est un TD mais aussi des éléments de cours à revoir et/ou à apprendre. Les séances se décomposeront alors généralement de la façon suivante :

- ½ h à ¾ h de rappel de cours selon les séances ;
- 1 petite pause de 5 à 10 minutes .
- Le reste de la séance consacré aux exercices d'application.

¤ Évaluation.

2 contrôles continus. Le premier aura lieu lors de **la 4ème séance et le deuxième lors de la dernière séance de TD.**

Attendus pour ces examens :

- Des exercices d'application à partir de ce qu'on aura vu en cours ;
- Interprétation et analyse des résultats statistiques obtenus ;
- Des questions de cours pour voir si vous avez bien compris et retenu ;

¤ Règles du cours.

- Présence obligatoire en TD. Règle du département. (**justificatifs à me présenter**)
- Au-delà de 5 min, les retardataires ne sont pas acceptés en cours.
- Pas de bavardages, pas de chahut => si je suis amené à demander à un étudiant trop dissipé de sortir, il ne sera plus considéré comme présent à cette séance ET/OU des points en moins sur la note de participation.
- J'attends de vous une participation active en cours : aussi bien pendant les moments où je vous laisserai travailler sur les exercices d'application que pendant les corrections => participation exigée (écrite et orale). N'hésitez pas à poser des questions s'il y a des choses que vous ne comprenez pas.

  Si vous avez des questions : vous pouvez les poser à tout moment en cours ou, si vous préférez, à la fin du cours.

C'est possible également par mail**.**

Vous pouvez aussi consulter deux livres, pas du tout jargonnants (pas des livres de maths) mais des livres de statistiques appliquées aux sciences sociales :

- **Fox W., *Statistiques sociales*, Paris, De Boeck, 1999**
- **Grenon G., Viau S., *Méthodes quantitatives en sciences humaines. Volume 1 : De l'échantillon vers la population*, Montréal, Gaëtan Morin Editeur, 1999**

- Si vous respectez ces règles et si vous vous impliquez dans le cours, c'est une bonne note assurée et facilement obtenue (ne serait-ce qu'avec la note de participation).

⇨ **Rassurer :** Pas un cours de maths, mais un cours de statistiques appliquées à la sociologie. Je ne suis pas mathématicienne mais chercheuse en sociologie ; je vous ferai travailler sur des choses utiles pour faire de la sociologie.

⇨ Le **but** du cours, ce n'est surtout pas de vous assommer avec plein de formules, avec des symboles, avec des annotations et des manières d'écrire compliquées. Je vous en donnerai le moins possible et j'essaierai toujours de dire et d'expliquer avec des mots ce qu'on cherche à faire avec telle ou telle formule : mettre en mots les formules plutôt qu'en symboles.

⇨ Des choses assez simples et basiques ; on ira de manière progressive de façon à ce que chacun puisse se mettre au niveau – en particulier pour ceux qui n'ont pas fait de maths depuis longtemps. Peut-être, du coup, ça paraîtra évident à certains, mais une bonne note assurée si vous jouez le jeu. On mettra l'accent également sur l'interprétation des données, des résultats, le sens qu'on peut leur donner en sociologie. D'ailleurs, dans la mesure du possible, je vous ferai travailler sur des exercices liés à des questions sociales. Le but n'est vraiment pas de vous coincer mais de vous faire acquérir les bases en statistiques et de vous montrer l'utilité de maîtriser ces outils.

Introduction, vocabulaire et éléments de base

Introduction

¤ L'importance de la quantification en sociologie : un outil d'objectivation du social.

La **statistique** : une discipline qui se développe à la fin du XIX^{e} siècle, en parallèle de la sociologie et sur laquelle s'appuie la sociologie. La statistique est un **outil très utilisé** en sociologie et ce, depuis la naissance de la discipline.

Le chiffre est un outil de connaissance du monde social. Durkheim s'en est servi dans *Le Suicide*. Dans cet ouvrage, il montre que le suicide, qui constitue un acte intime, n'obéit pas seulement à des causes individuelles / psychologiques. Il ne se produit pas par hasard. Il relève certaines régularités : le taux de suicide varie géographiquement (selon les sociétés) et, dans chaque société, il est plutôt stable. Les hommes se suicident plus que les femmes, les non croyants plus que les croyants, etc. Outre les motifs personnels, les variations du taux de suicide obéissent donc à une loi sociale. De son point de vue, la statistique constitue un outil pour **objectiver le social**, pour se détacher des prénotions. Un outil d'objectivation du social, des phénomènes sociaux (mobilité sociale, inégalités hommes-femmes).

C'est un **outil d'administration de la preuve** : énoncer un fait avec une statistique à l'appui cela paraît plus probant que si ce même fait est énoncé à partir d'observations ou à partir d'un corpus de quelques dizaines d'entretiens. Les méthodes quantitatives ont leurs défauts et leurs inconvénients mais un résultat semble souvent moins discutable s'il est fondé sur des statistiques que s'il est fondé sur des entretiens, qui paraissent moins crédibles et trop subjectifs. Ça, c'est une question épistémologique.

⇨ **Savoir manipuler les chiffres est une nécessité**.

¤ Qu'entend-on par statistique ?

Dans la vie courante, une statistique désigne un nombre qui résume un ensemble de données sur une population, qui synthétise des informations pour les rendre plus intelligibles. Elle réduit l'information en la rendant plus compréhensible : il est difficile de déceler une cohérence dans les 1500 votes lors d'une élection sans résumer ceux-ci. Pour donner un autre exemple de données quantitatives qui synthétisent des informations sur la population : en 2012, les femmes se marient en moyenne à 30 ans, les hommes à 32 ans ; trente ans plus tôt, les femmes avaient en moyenne 23 ans au moment de leur mariage et les hommes avaient 25 ans. Ces résumés quantitatifs sont pratiques car ils informent rapidement d'une évolution (ici, celle de l'âge au

moment du mariage, qui a augmenté en trente ans). On a besoin de ces résumés quantitatifs qui condensent l'information. Des institutions statistiques (lesquelles ?).
Le terme statistique renvoie également à la méthode utilisée pour calculer ces résumés numériques : les pourcentages (calculer la proportion de filles et de garçons dans le groupe), les moyennes (l'âge moyen des étudiants du groupe) ou des représentations graphiques qui, sous la forme de courbes ou de diagrammes, permettent de présenter l'information autrement.
En revanche, si la statistique permet de mettre en évidence des phénomènes, elle ne fournit en aucun cas d'explication. L'explication, l'interprétation et l'analyse de ces phénomènes est l'affaire du sociologue. C'est un outil d'aide précieux pour l'analyse, qu'il est nécessaire de connaître mais il ne remplace pas l'analyse.

¤ Statistique univariée. On va faire de la statistique descriptive, plus précisément univariée. On va travailler sur une variable à la fois. En L2, croisement de deux variables (voir si le sexe influence le niveau de diplôme ou le niveau de revenu par exemple) => statistique bi-variée.

I. Vocabulaire

L'individu (ou unité statistique) : l'analyse statistique démarre par l'identification du groupe soumis à l'étude. On sait qui sont les individus en répondant à la question : « **sur qui porte l'étude ?** » Il peut s'agir d'humains, d'animaux ou d'objets. Exemples :

- les étudiants en 1ère année de sociologie inscrits à l'université de Tours en 2016/2017 ;
- les ménages en France en 2011 ;
- les bovins élevés dans l'espace de l'UE en 2010.

En socio, les individus sont le plus souvent soit des personnes, soit des ménages.

La population : la population est constituée de la totalité des individus sur lequel porte l'étude. La population est donc l'ensemble des individus visés par l'étude. À chaque recherche correspond une population précise. C'est un ensemble fini. Son effectif est aussi appelé **taille** et il est **noté N**. Chaque élément de la population est identifié par un numéro compris entre 1 et N (numéro arbitraire). Un numéro n'est porté que par un seul individu.
Exemple : population = groupe de TD ; sa taille = … ; l'unité statistique : étudiant.

Variables : les variables sont des propriétés qui permettent de décrire les individus de la population. On arrive à trouver quelles sont les variables en répondant à la question : « **sur quoi porte l'étude ?** » Une variable est donc une caractéristique dont la valeur peut différer d'un individu à l'autre.

Exemple : en considérant les étudiants inscrits en sociologie à l'université de Tours, on peut les caractériser par leur âge, leur sexe, la catégorie socioprofessionnelle de leurs parents, le nombre de frères et sœurs, le type de logement qu'ils occupent.

L'âge, le sexe, la CSP des parents, la fratrie, le logement, les revenus sont des variables.

Modalités d'une variable : c'est l'ensemble des valeurs possibles de la variable. Les modalités sont les différentes réponses que l'on peut trouver dans l'ensemble des données.

Exemples : la variable statut matrimonial peut avoir pour modalités : marié, en couple, pacsé, célibataire, veuf, divorcé. La variable sexe a pour modalité : homme ou femme.

II. Les données construites et les tableaux de distribution

Une fois qu'on a recueilli les données, on les réorganise et on va procéder à des comptages. Pour les organiser, on fait un **tri à plat**, en comptant le nombre d'individu présentant chaque modalité de la variable. Ce faisant, on est en fait en train de calculer les effectifs.

L'effectif : c'est le nombre de réponses associées à chaque modalité de la variable. L'effectif, c'est une manière simple de résumer des informations : on compte le nombre de cas pour chaque modalité de la variable. On ne note n_i et on lit n indice i ou ni

Lorsqu'on additionne le total des effectifs, on doit retomber sur **N** (taille de la population).

Structure des familles avec enfant(s) de moins de 18 ans en 1990 et en 2010

Structures	Familles (en milliers)	
	1990	2010
Couples avec enfant(s)	6 699,5	6 257, 4
Familles monoparentales	952,7	1 686,7
Ensemble	7 652,2	7 944,1

En 1990 : -le premier effectif n_1= 6 699,5 c'est celui des couples avec enfant(s)

-le deuxième effectif n_2 = 952,7 c'est celui des familles monoparentales.

-L'effectif total en 1990 est : N = 7 652,2

Les pourcentages

Les **effectifs** sont utiles pour résumer l'information, en revanche, **ils sont difficiles à interpréter**. La comparaison d'effectifs se complique s'ils ne sont pas basés sur le même nombre de cas. Que pouvons-nous dire de l'évolution de la structure des familles entre 1990 et

2010 ? Les couples avec enfants sont-ils moins fréquents en 1990 qu'en 2010 ? Il est difficile de comparer la composition des ménages puisque le nombre de familles a augmenté.

Pour comparer deux distributions, il faut **standardiser les données**. Pour ce faire, on calcule quel serait chacun des effectifs si le nombre total de cas était exactement 100. On appelle **pourcentage**, le résultat de cette standardisation.

Structure des familles avec enfant(s) de moins de 18 ans en 1990 et en 2010

Structures	Pourcentages	
	1990	2010
Couples avec enfant(s)	87,6	78,8
Familles monoparentales	12,4	21,2
Ensemble	100,0	100,0

On est très familier avec les pourcentages, de sorte qu'on songe rarement à ce qu'ils signifient réellement : les pourcentages sont ce que les effectifs seraient s'il y avait 100 cas au total. On voit dans le cas du nombre de familles avec enfants, que la proportion de couples a sensiblement diminué entre 1990 et 2010, alors que le nombre de familles monoparentales a fortement augmenté, notamment chez les femmes.

Si on reprend formellement : **le pourcentage** indique, sur une base de 100, quelle partie de la population correspond à la modalité étudiée. On l'obtient par un calcul simple : une division et une multiplication. On divise l'effectif par le nombre total d'unités la population puis on multiplie le résultat par 100. L'addition d'un ensemble de pourcentages doit retomber sur 100.

$$\text{Pourcentage} = \frac{effectif}{N} * 100$$

Exemple : Dans une classe, il y a 25 filles et 10 garçons. Déterminons le pourcentage des filles :

N=25+10 = 35

En appliquant la formule : **Pourcentage =** $\frac{effectif}{N} * 100$ **:**

le pourcentage des filles est = $\frac{25}{35}$ x 100 = 0,714 x 100 = 71,4

Les filles représentent 71,4% de la population de la classe.

¤ **Tableau statistique**.

Titre du tableau.

Variable	Effectifs	Pourcentages
Modalités de la variable	nb de cas	

Modalités de la variable		
Total	N	100

Source du tableau.

☝ Le symbole % n'apparaît pas dans le tableau. Ils sont superflus et encombrent le tableau.

¤ **Taux de variation.**

Un outil pour **caractériser et quantifier** une évolution d'une valeur par rapport à sa valeur de départ. Il mesure la variation d'un phénomène entre deux dates. Plus concrètement : on connaît deux grandeurs (qu'on peut appeler valeur de départ et valeur d'arrivée) et on se demande ce que vaut la différence entre les deux. Cet écart entre deux grandeurs porte le nom de taux de variation. Cette variation, on l'exprime souvent en pourcentage. On l'obtient par un calcul simple : une soustraction, une division et une multiplication.

La formule du taux de variation est

$$\text{Taux de variation : } \frac{valeur\ d'arrivée - valeur\ de\ départ}{valeur\ de\ départ} \times 100$$

Exemple. On cherche à estimer la variation des naissances en France entre 2005 et 2006, à partir du tableau suivant.

Nombre de naissances vivantes en France (Insee, 2008-2009)

2005	2006	2007	2008
806 000	829 300	818 700	834 000

Taux de variation du nombre de naissances entre 2005 et 2006 :

$$\frac{Valeur\ d'arrivée - valeur\ de\ départ}{Valeur\ de\ départ} \times 100 = \frac{829300 - 806000}{806000} \times 100 = \frac{23300}{806000} \times 100 = 2{,}89$$

La valeur est positive, il s'agit donc d'une augmentation. On dira alors : en France, entre 2005 et 2006, le nombre de naissances a augmenté / a crû de 2,89 %.

Interprétation du taux de variation.

- Quand il est strictement positif, on dit que la valeur de la variable augmente ;
- Quand il est strictement négatif, on dit que la valeur de la variable diminue ;
- Quand il est égal à zéro, la valeur de la variable reste inchangée.

Le sens de la variation / de l'évolution peut augmenter ou baisser. Il n'augmente pas toujours. Il est impératif de préciser la période, le lieu, la variable étudiée et le sens de la variation.

Attention à ne pas confondre frein et diminution : lorsqu'un taux de variation diminue au cours du temps, cela ne signifie pas que la variable diminue (il faudrait que le taux de variation soit négatif), mais qu'elle augmente de moins en moins vite. A l'inverse, lorsqu'un taux de variation augmente au cours du temps, cela signifie que la variable augmente de plus en plus vite. Ainsi, le nombre de naissances a été de 834 000 en 2008, il y a plus de naissances qu'en 2007, même si le taux de croissance du nombre de naissances entre 2007 et 2008 est plus faible qu'entre 2005 et 2006. La croissance ralentit mais le nombre de naissances continue d'augmenter. Elle augmente mais moins vite.

Comparer deux taux de variation : il suffit de les soustraire l'un à l'autre. **La différence obtenue s'exprime en points** (sous-entendu en points de pourcentage), et non pas en %.

Attention. Les valeurs d'origine et d'arrivée doivent être proportionnelles et porter sur le même objet. Ça n'a pas de sens de le calculer entre un salaire et la vitesse de fonte des glaces.

Séance 2. Les différents types de variables et leur échelle de mesure

Différents types de variables

Il y a différents types de variables. Connaître et repérer le type de variables auxquelles on a affaire est essentiel car cela détermine ce qu'on peut faire d'un point de vue statistique avec chacune d'entre elles, les calculs qu'on peut opérer.

¤ Les variables qualitatives.

Une variable est dite qualitative si les modalités de cette variable sont **des mots ou des expressions qui ne correspondent pas à des quantités numériques**. Ses modalités ne sont **ni mesurables, ni quantifiables**. Par exemple : les CSP, le sexe, la langue maternelle, la situation matrimoniale, la nationalité, la commune de naissance, etc.

¤ Les variables quantitatives.

Une variable est dite quantitative si les modalités de cette variable sont des **quantités numériques**. Ses modalités sont **mesurables et quantifiables**.

On y distingue les variables quantitatives discrètes ou discontinues d'une part et les variables quantitatives continues de l'autre.

- **Les variables quantitatives discrètes ou discontinues** sont celles qui ne peuvent prendre que des **valeurs entières**. La taille d'une famille est un exemple de variable discrète : une famille peut avoir 2, 3 ou 4 membres, mais ne peut pas avoir 1,7 membres.

-**Une variable quantitative est dite continue** quand les données recueillies prennent **aussi bien des valeurs entières que des décimales**. Souvent et en pratique, ce sont des variables dont le nombre de modalités qu'elles peuvent prendre est si grand qu'on est obligé de procéder à un regroupement par classes pour leur description et leur traitement.

Ex. pour la variable âge, il peut y avoir une centaine de modalités de réponse => classes d'âge, fourchettes de revenu, etc. Les valeurs possibles de la variable sont donc *a priori* en nombre infini et elles s'expriment dans des continuums qui peuvent être segmentés.

- **Regroupement en classes des données**.

Pour constituer les classes, on construit des intervalles de valeur :

Une classe est un intervalle semi ouvert à droite :

[a ; b[

Exemple : pour la variable âge on peut avoir comme classes :

[0 ; 5[[5 ; 15[[15 ; 20[[20 ;30[etc

Amplitude de classe

On peut calculer pour chaque classe, son **amplitude** (notée a_i**).**

Pour la classe [a ; b [son amplitude est : $a_i = b-a$

Exemple : Pour [20 ; 30[son amplitude est : $a_i = 30 - 20 = 10$

Pour la classe [0 ;5|, son amplitude est 5 - 0 = 5 etc.

Centre de la classe

Le centre de classe (noté c_i ou x_i) est la demi somme des 2 bornes de la classe

Pour la classe **[a ; b [son centre de classe est :** $\mathbf{c_i} = \frac{a+b}{2}$

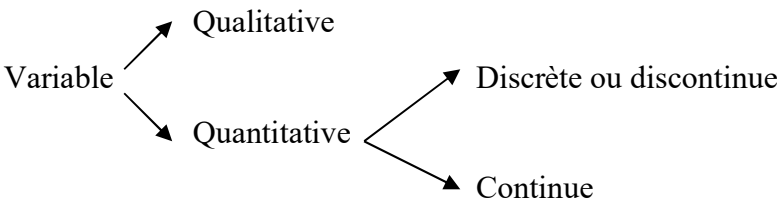
Pour [0 ; 25[$\mathbf{c_i} = \frac{0+25}{2} = 12{,}5$

Pour [25 ; 35[$\mathbf{c_i} = \frac{25+35}{2} = 30$; etc.

Pour les calculs, on se servira du centre de la classe.

Au total : on peut faire apparaître le **schéma** suivant :

Types de variables statistiques



Les échelles de mesure des variables statistiques

Les variables peuvent être classées sous une autre forme, selon la façon dont elles se mesurent : on peut parler de variable nominale, ordinale, métrique. Connaître à quelle échelle de mesure

renvoie telle variable est là aussi nécessaire pour savoir quels sont les traitements statistiques qui sont autorisés avec elle, c'est-à-dire s'ils ont du sens.

¤ **L'échelle nominale.**

Les échelles nominales correspondent à des variables qualitatives (et discrètes). La répartition d'une population en catégories socioprofessionnelles (CSP) constitue un cas typique d'échelle nominale. Une relation d'équivalence est définie entre les différentes modalités de la variable : elles sont considérées comme égales les unes entre elles et il n'y a pas d'ordre entre elles. Les modalités n'ont pas de relation d'ordre entre elles. Autre exemple, le sexe.

On peut assigner – pour faciliter le traitement informatique des données par exemple – un numéro à chaque modalité de la variable, mais c'est purement conventionnel. On peut par exemple accoler un « 1 » à masculin et un « 2 » à féminin mais ces nombres ne sont ici que des symboles arbitraires. On peut inverser les numéros donnés. Ainsi, si on désigne dans une enquête la situation familiale par : « 1 » célibataire, « 2 » marié, « 3 » veuf, « 4 » divorcé, « 5 » pacsé, cela n'implique évidemment pas que la situation « 3 » soit supérieure à la situation « 2 ». Les variables nominales sont simplement des classifications et leurs valeurs ne peuvent être ordonnées. Ils n'impliquent aucun ordre ou rang.

Avec ce type d'échelle, les opérations statistiques sont assez limitées :

- on peut compter le nombre d'individus (effectifs et pourcentages) ;
- on peut rechercher le mode, qu'on étudiera dans quelques séances ;
- on peut construire des tableaux.

Le calcul d'une moyenne des codes de situation familiale ou de CSP n'a pas de sens, pas plus que le calcul d'une moyenne des codes départementaux d'une pile de lettres !

¤ **L'échelle ordinale.**

Une variable ordinale (qui s'apparente également aux variables qualitatives discrètes) présente les mêmes caractéristiques que l'échelle précédente, mais avec une relation d'ordre entre les modalités de cette variable. C'est une variable dont les modalités peuvent être ordonnées. Les modalités ont une relation d'ordre entre elles ; plus précisément : elles sont considérées comme égales, mais l'une sera tenue supérieure / plus grande à une autre.

Un exemple de variable ordinale, les échelles d'opinion. Quand on demande aux personnes si elles sont « pas du tout d'accord », « plutôt pas d'accord », « plutôt d'accord », « entièrement

d'accord » avec une quelconque opinion. Il y a un ordre sous-jacent. On peut assigner à ces valeurs des nombres qui indiquent leur ordre. On peut ordonner des classes de revenus, comme (revenus faibles, moyens, élevés). De même avec le niveau de diplôme.

Les variables ordinales fournissent donc une information supplémentaire par rapport aux variables nominales : les variables ordinales permettent en plus de disposer les cas sur un continuum (sur une échelle). Comme pour les variables nominales, il n'est pas possible de faire des opérations arithmétiques avec les variables ordinales. On peut faire les mêmes opérations qu'avec les variables nominales, mais en plus, on peut avec les variables ordinales déterminer la médiane (on y reviendra dans quelques séances).

¤ **L'échelle métrique.**

L'échelle métrique correspond aux variables quantitatives. Je ne rentrerai pas plus dans le détail pour cette échelle. La plupart des opérations statistiques sont autorisées.

Tableau récapitulatif.

Type d'échelle de mesure	**Relations entre les modalités**	**Opérations statistiques**	**Caractéristiques**	**Exemples**
« Échelle » nominale	Aucun ordre.	Mode. Effectifs et %. Tableaux.	Qualitative.	CSP Sexe Situation matrimoniale
Échelle ordinale	Ordre.	Mode. Effectifs et %. Médiane, quartiles, déciles.	Qualitative.	Préférences. Attitudes.
Échelle métrique	Ordre. Rapport entre deux intervalles.	La plupart des opérations sont autorisées.	Quantitative. Discrète ou continue.	Âge. Prix et revenus.

Séance 3. Construire des tableaux statistiques et représentations graphiques

Présenter les données sous forme de tableaux et de graphiques, cela permet de donner lisibilité et intelligibilité aux données. On les présente de manière plus parlante.
On ne procède pas de la même façon selon les variables.

I. Les tableaux

A) Les variables qualitatives à échelle nominale.

- Étape 1 : Titrer le tableau. Il faut toujours titrer un tableau. Celui-ci doit informer clairement le lecteur sur le contenu.

  Une formulation de titre (en l'adaptant à chaque situation) : **répartition des individus en fonction de la variable, à telle période**.
- Étape 2 : Dans la première colonne du tableau, on inscrit la variable et toutes ses modalités. Le titre de cette colonne correspond au nom de la variable.
- Étape 3 : Dans la 2e colonne, on indique le nombre d'individus pour chaque modalité (l'effectif).
- Étape 4 : Dans la 3e colonne, on donne les pourcentages pour chaque modalité.
  (les pourcentages cumulés n'ont pas de sens, car il n'y a pas de relation d'ordre)

Titre : Répartition des catégories socioprofessionnelles, en 2013, d'une entreprise française.

Catégories socioprof	Effectifs	Pourcentages
Cadres	64	10
Techniciens	160	25
Employés	128	20
Ouvriers	192	30
Personnel de service	96	15
Total	640	100

B) Variables qualitatives à échelle ordinale

Quasiment la même chose :

- Étape 1 : Titrer le tableau.

- Étape 2 : Indiquer les modalités de la variable, dans la première colonne du tableau. Ce qui change : les modalités sont placées en ordre croissant.
- Étape 3 : Préciser le nombre d'individus pour chaque modalité (colonne des effectifs).
- Étape 4 : Déterminer le pourcentage des individus (3ème colonne).
- Étape 5 : Indiquer, dans une 4ème colonne, les pourcentages cumulés croissants (pour déterminer graphiquement la médiane).

Titre : répartition des parents selon le niveau d'importance accordée au sport à l'école

Niveau de l'importance accordée	Nombre de parents	Pourcentage	Pourcentages cumulés
Pas du tout	33	3,27	3,27
Peu	76	7,53	10,80
Assez	193	19,13	29,93
Beaucoup	707	70,07	100,00
Total	1009	100,00	

C) Variables quantitatives discrètes

Même chose que dans le cas des variables qualitatives à échelle ordinale.

Titre : répartition du nombre d'enfants par famille, en 2008, d'une commune française

Nombre d'enfants de la famille	Effectif	%	% cumulés
0	442	11,5	11,5
1	796	20,7	32,2
2	981	25,5	57,7
3	689	17,9	75,7
4	398	10,4	86,0
5	231	6,0	92,0
6	126	3,3	95,3
7	70	1,8	97,1
8	43	1,1	98,3
9	29	0,8	99,0
10	19	0,5	99,5
11	12	0,3	99,8
12	7	0,2	100,0
Total	3843	100,0	

D) Variables quantitatives continues

Là, les choses changent un peu car on peut avoir un très grand nombre de modalités (sauf si le choix de réponses se fait déjà sous forme de classes).

- Étape 1 : Titrer le tableau.
- Étape 2 : Délimiter les classes dans la première colonne.
    - D'abord, on détermine le nombre de classes le plus approprié pour grouper les données. Pas un choix si facile et si évident. Il faut opter pour un nombre de classes qui ne soit ni trop petit ni trop grand, que les classes formées soient pertinentes et que cela ait du sens d'un point de vue sociologique.
    - Évaluer l'étendue des données. Pour cela, on estime l'écart entre la plus petite et la plus grande donnée et le reste de cette soustraction donne l'étendue des données.
    - Déterminer la largeur des classes. On peut élaborer des classes de largeurs égales mais ça n'est pas toujours le cas. Là, encore, ce qui gouverne le choix : la pertinence sociologique. Par exemple, pour l'âge on préférera [0-18[[18-30[à [0-15[[15-30[
    - Former les classes. Il s'agit de choisir le point de départ de la première classe, c'est-à-dire la borne inférieure de la première classe à partir de laquelle les autres seront déterminées. Chaque valeur des données doit entrer dans une classe (exhaustivité) et une seule (exclusivité). Pour respecter ces principes, une convention : la borne inférieure est incluse dans la classe tandis que la borne supérieure en est exclue (« De … à moins de… »).
- Étape 3 : Déterminer le point milieu. On détermine le centre de la classe qu'on peut indiquer dans le tableau, dans la 2ème colonne par exemple. Le centre de la classe s'obtient en ajoutant les valeurs de la borne inférieure et de la borne supérieure en divisant cette somme par 2. Le centre de la classe permettra de tracer les graphiques et de calculer la moyenne et l'écart-type.
- Étape 4 : Indiquer le nombre d'individus par classe (colonne des effectifs).
- Étape 5 : Déterminer le pourcentage des individus par classe.
- Étape 6 : Indiquer les pourcentages cumulés croissants.

Répartition de 175 ménages français, en 2012, en fonction des coûts mensuels d'habitation.

Coûts mensuel d'habitation	**Centre de la classe**	**Nombre de ménages**	**% des ménages**	**% cumulé des ménages**
[200 ; 300[	250	11	6.29	6.29
[300 ; 400[	350	13	7.43	13.72
[400 ; 500[	450	13	7.43	21.15
[500 ; 600[	550	14	8.00	29.15
[600 ; 700 [	650	16	9.14	38.29

[700 ; 800[	750	15	8.57	46.86
[800 ; 900[	850	26	14.86	61.72
[900 ; 1000[	950	35	20.00	81.72
[1000 ; 1100[	1050	32	18.29	100.00
Total		175	100.00	

Trois derniers éléments à propos des tableaux :

- par convention, on donne toujours le même nombre de décimales (toujours deux ou toujours une, même si c'est pour finir par un zéro) ;
- par convention, le symbole % n'apparaît pas dans le tableau (à la limite dans la ligne de titre). Il est superflu et il encombre le tableau ;
- on aligne, dans la mesure du possible, les nombres.

Concernant la **question des arrondis**.
On arrondit à la décimale supérieure lorsque le chiffre suivant est compris entre 6 et 9. On garde la même décimale si le chiffre suivant est compris entre 0 et 4. Si la décimale suivante est un 5, pour savoir si on doit arrondir à la décimale supérieure, il suffit de regarder après le 5 et si le chiffre est suivant est compris entre 0 et 4, on n'arrondit pas à la décimale supérieure mais si le chiffre après le 5 est compris entre 6 et 9, on arrondit à la décimale supérieure.
Par exemple : on aura 10,45 (si la décimale après le 5 est comprise entre 0 et 4) mais 10,46 (si la décimale après le 5 est comprise entre 6 et 9). Si la décimale d'après est 5, pour déterminer l'arrondi, on regarde la décimale placée encore à la suite du 5.

II. Les représentations graphiques

On va étudier deux types de représentations graphiques : les diagrammes (dont la construction varie selon le type de variable) et des courbes. Ces représentations graphiques se construisent à partir des **pourcentages** (et non des effectifs).

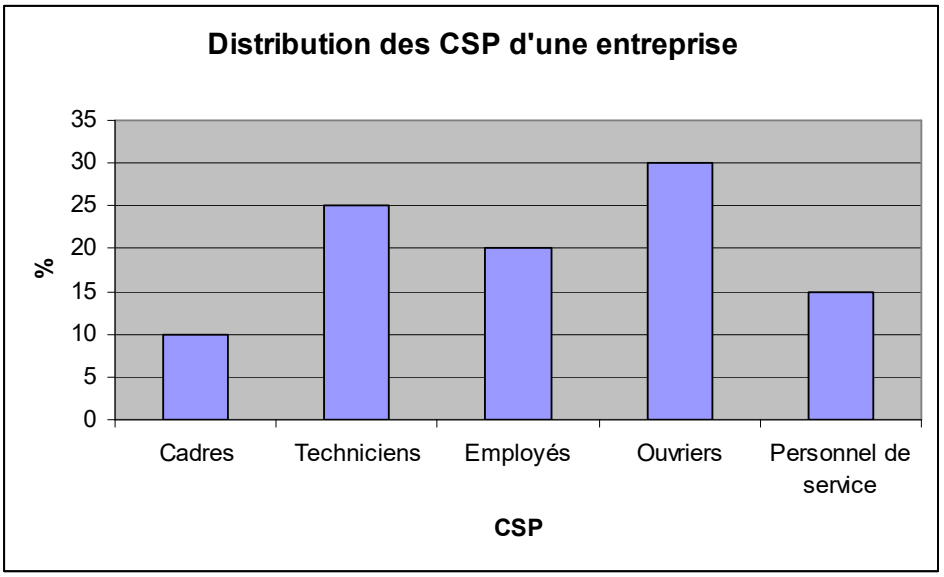
A) Les variables qualitatives à échelle nominale et à échelle ordinale

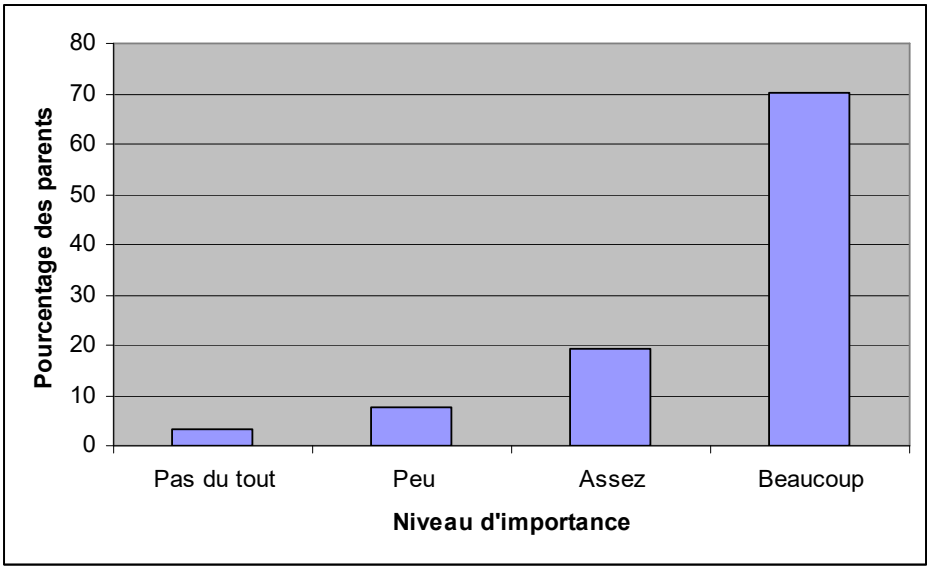
=> le diagramme à bandes verticales et le diagramme linéaire.

Le diagramme à bandes verticales

- Étape 1 : Titrer la figure. Il peut être le même que celui du tableau correspondant, puisqu'il représente la même répartition mais sous une forme visuelle.

- Étape 2 : Placer les modalités de la variable sur l'axe horizontal des abscisses. La distance entre deux valeurs doit toujours être la même. Pour les variables qualitatives à échelle ordinale, il faut placer les modalités de la variable en ordre croissant (ou décroissant).
- Étape 3 : Placer les pourcentages des unités sur l'axe vertical des ordonnées.
- Étape 4 : Tracer les bandes, dont la hauteur correspond au pourcentage lié à chaque modalité de la variable. Les bandes sont toutes de même largeur.

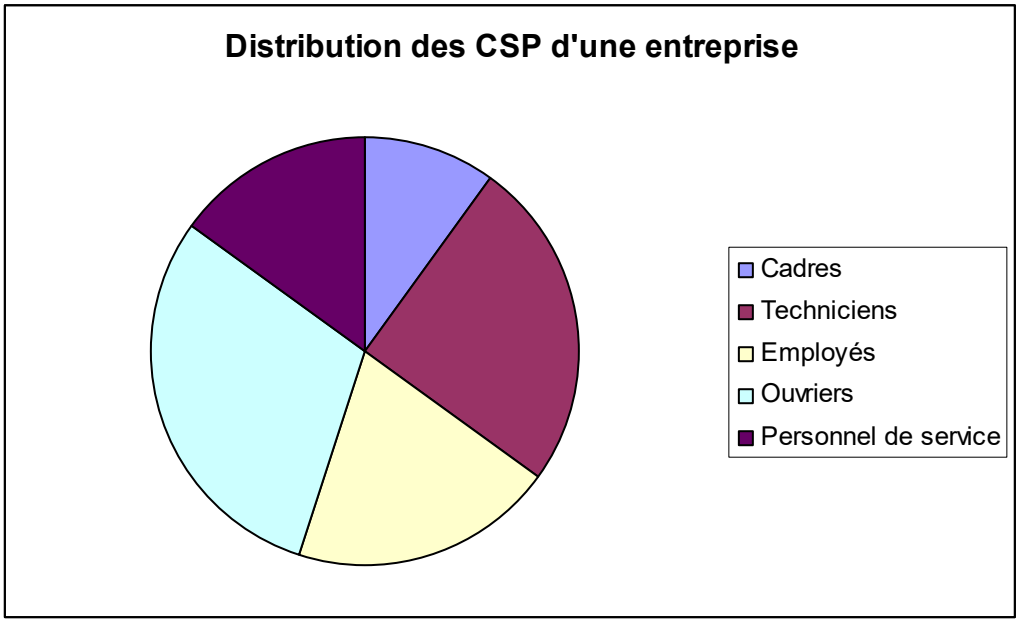




Le diagramme circulaire

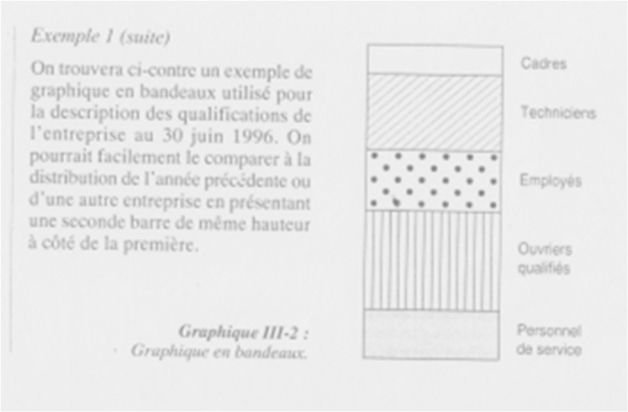
- Étape 1 : Titrer la figure.
- Étape 2 : Convertir les pourcentages en degrés (on multiplie le pourcentage par 360).
- Étape 3 : Déterminer les secteurs.

- À éviter si nombre trop important de modalités => question de lisibilité.



Le diagramme linéaire (ou le diagramme en bandeaux)

- Étape 1 : Titrer la figure.
- Étape 2 : Établir l'axe horizontal. L'axe horizontal est établi tout simplement en fonction d'une échelle qui va de 0 % à 100 %.
- Étape 3 : Construire les bandes. Il s'agit de placer bout à bout les bandes, une par modalités. Chaque bande a une longueur égale au pourcentage de la modalité de la variable.


Exemple 1 (suite)

On trouvera ci-contre un exemple de graphique en bandeaux utilisé pour la description des qualifications de l'entreprise au 30 juin 1996. On pourrait facilement le comparer à la distribution de l'année précédente ou d'une autre entreprise en présentant une seconde barre de même hauteur à côté de la première.

Graphique III-2 : Graphique en bandeaux.

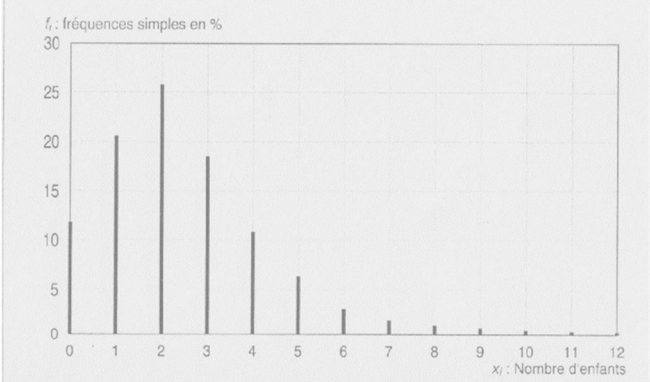
C) <u>Les variables quantitatives discrètes</u>.

Par convention, on représente les variables quantitatives discrètes par un diagramme en bâtons.

- Étape 1 : Titrer le graphique.

- Étape 2 : Placer les valeurs de la variable sur l'axe horizontal. On place les modalités de la variable sur l'axe de l'abscisse, de la plus petite à la plus grande valeur. La distance entre deux valeurs doit toujours être la même. Cet axe est désigné par le nom de la variable et, s'il y a lieu, il faut préciser les unités de mesure utilisées.
- Étape 3 : Placer les pourcentages des unités statistiques sur l'axe des ordonnées. On établit une échelle pour placer les pourcentages. Si le pourcentage le plus élevé est – imaginons – 58,42%, il est inutile d'aller plus haut que 60%. Cette échelle permet d'estimer la hauteur des bâtons. Il faut également titrer cet axe.
- Étape 4 : Tracer les bâtons. Pour chaque modalité de la variable, sur l'axe horizontal, on trace un bâton dont la hauteur correspond au pourcentage qui lui correspond. Puisque la variable est discrète, les bâtons doivent être espacés et très minces pour illustrer le fait que chaque bâton est associé à une seule valeur de la variable.

Au-dessus de l'abscisse, on élève un bâtonnet de hauteur égale ou proportionnelle à l'effectif.



Graphique III-3 : Diagramme en bâtons de la distribution des familles.

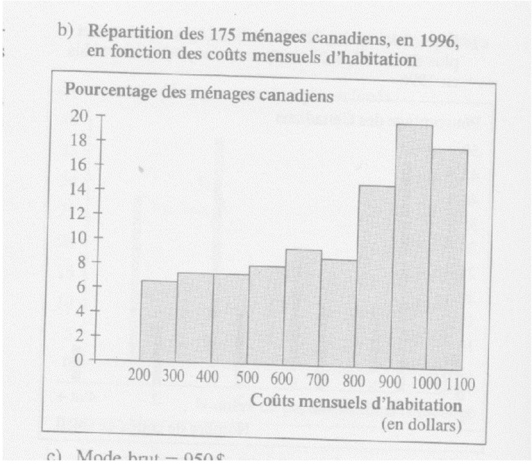
D) Les variables quantitatives continues.

Dans ce cas, on utilise l'histogramme (appellation spécifique à ces variables) ou le polygone des pourcentages ou la courbe des pourcentages cumulés.

L'histogramme

- Étape 1 : Titrer l'histogramme.
- Étape 2 : Placer les valeurs de la variable (axe horizontal des abscisses). On situe les bornes des classes en ordre croissant vers la droite, on indique son nom et les unités de mesure. Ne pas laisser d'espace entre les classes (puisque caractère continu).
- Étape 3 : Placer les pourcentages des individus sur l'ordonnée. Sur l'axe vertical, on trace une échelle pour les pourcentages.

- Étape 4 : Pour chaque classe, on trace un rectangle dont la base est la largeur de la classe et la hauteur, le pourcentage des classes. **Les rectangles sont collés**.

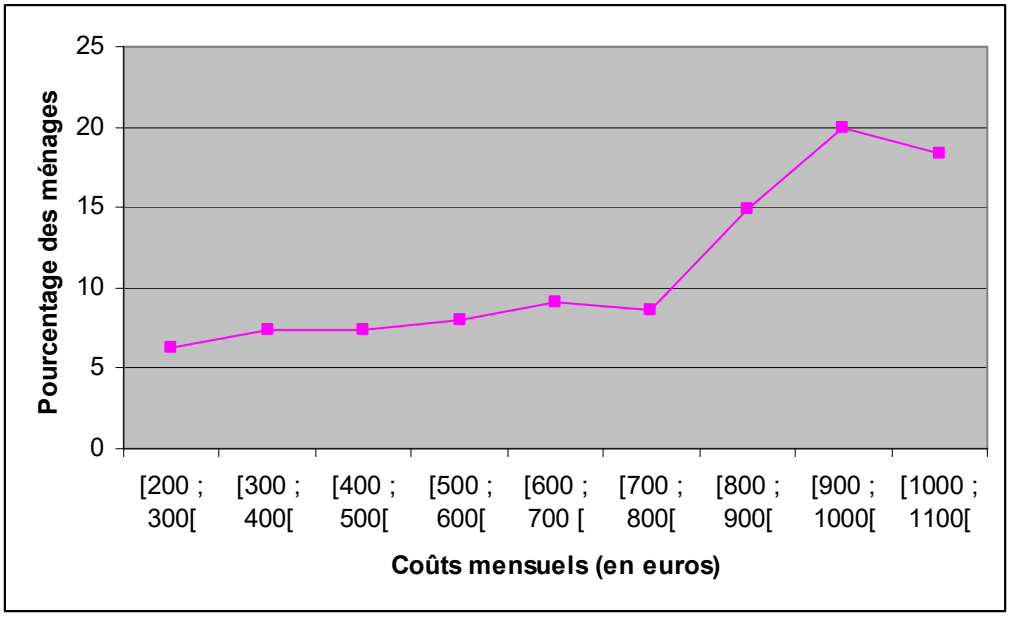


Le polygone des pourcentages

Le polygone donne l'allure générale de la distribution. On peut placer plusieurs polygones sur un même graphique (impossible avec un histogramme). Ce qui facilite les comparaisons.

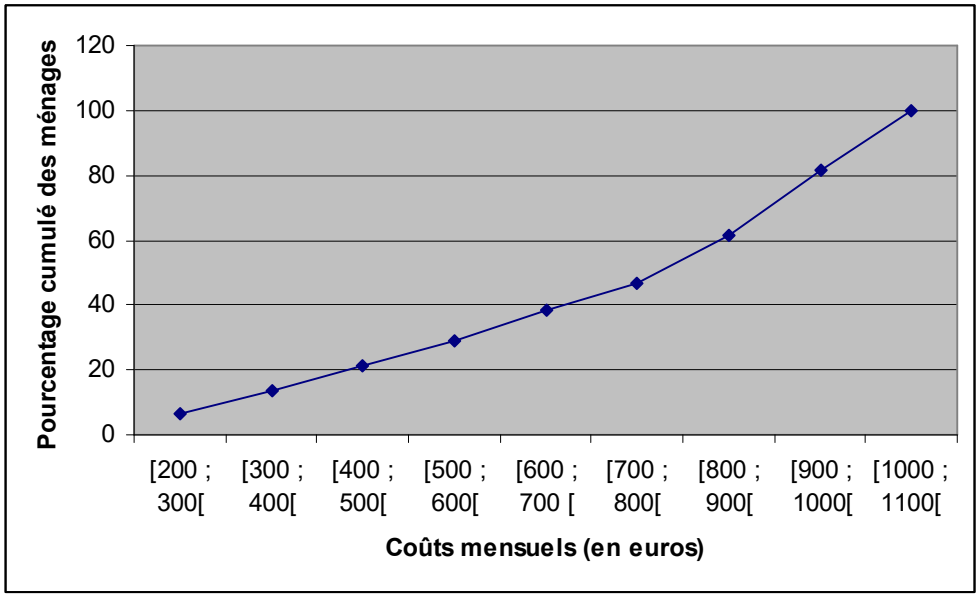
- Étape 1 : Titrer la figure.
- Étape 2 : Placer les valeurs de la variable sur l'axe horizontal.
- Étape 3 : Placer les pourcentages des unités statistiques sur l'axe vertical. Pour chaque classe, on place un point à la hauteur égale au pourcentage des unités dans la classe.
- Étape 4 : Tracer les segments en reliant les points de l'étape 3 entre eux.

Titre : Répartition des 175 ménages, en 2012, en fonction des coûts mensuels d'habitation



La courbe des pourcentages cumulés

Principe similaire au polygone des pourcentages, en se basant sur les pourcentages cumulés croissants.



☝ Cas particulier : le cas de classes de largeurs inégales.

Titre du tableau : Répartition des salaires annuels d'une grande entreprise, en 2007

Salaire (en K€)	Effectifs	Centre de la classe	Pourcentages	% cumulés
moins de 15	390	7,5	3,07	3,07
15 à 20	230	17,5	1,81	4,88
20 à 25	454	22,5	3,57	8,45
25 à 30	1080	27,5	8,5	16,95
30 à 35	1420	32,5	11,18	28,13
35 à 40	1565	37,5	12,32	40,45
40 à 50	2737	45	21,54	61,99
50 à 60	1746	55	13,74	75,73
60 à 70	1009	65	7,94	83,67
70 à 80	598	75	4,71	88,38
80 à 100	631	90	4,97	93,35
100 et plus	848	(250)	6,65	100,00
Total	12 705		100,00	

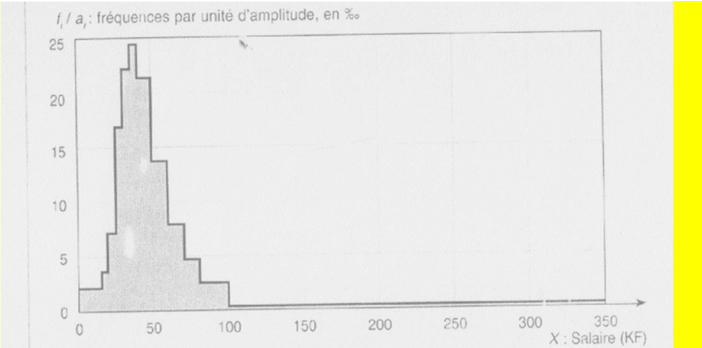
Faute d'informations, on suppose que le centre de la dernière classe est 250.

Le principe d'un histogramme est que l'aire de chacun de ces rectangles doit être proportionnelle au pourcentage des données dans la classe, ce qui permet de déterminer l'importance relative d'une classe par rapport à une autre. Dans le cas des classes égales, cette

propriété est respectée. Mais ce n'est pas le cas lorsqu'une classe au moins a une largeur différente de celle des autres. Pour respecter cette propriété, la surface du rectangle doit être égale au pourcentage des données dans la classe, de telle sorte que la somme de toutes les surfaces donne 100 %. Pour y arriver, il faut construire des rectangles dont la hauteur est égale au pourcentage divisé par la largeur de la classe : hauteur = pourcentage / largeur. Et, la largeur, c'est l'amplitude de la classe.

On peut donc ajouter deux colonnes à ce tableau : l'amplitude de chaque classe et la densité.

Salaire (en K€)	Effectifs	Centre de la classe	Amplitude de la classe	Pourcentages	Densité des salaires	% cumulés
- de 15	390	7,5	15	3,07	0,205	3,07
15 à 20	230	17,5	5	1,81	0,362	4,88
20 à 25	454	22,5	5	3,57	0,714	8,45
25 à 30	1080	27,5	5	8,5	1,700	16,95
30 à 35	1420	32,5	5	11,18	2,236	28,13
35 à 40	1565	37,5	5	12,32	2,464	40,45
40 à 50	2737	45	10	21,54	2,154	61,99
50 à 60	1746	55	10	13,74	1,374	75,73
60 à 70	1009	65	10	7,94	0,794	83,67
70 à 80	598	75	10	4,71	0,471	88,38
80 à 100	631	90	20	4,97	0,248	93,35
100 et +	848	(250)	(250)	6,65	0,084	100,00
Total	12 705			100,00		



Graphique III-7 : Histogramme de la distribution des salaires.

)=

Séance 6. Les indicateurs de position

Les indicateurs de position, appelés aussi mesures de tendances centrales, ont pour but de résumer les informations que contiennent une distribution statistique univariée. Ce sont des mesures qui permettent de **se faire une idée de la forme de la population**, de **la manière dont elle se répartit.**

On va étudier trois indicateurs de position : le mode, la moyenne et la médiane.

I. Le mode

Le mode, c'est la modalité de la variable qui a le plus grand nombre d'individus, c'est-à-dire qui a **le pourcentage le plus élevé**.

- ⇨ Un ex. : en France, il y a plus de femmes que d'hommes. On dit alors que le sexe modal est le féminin.

Il n'existe **aucune formule** qui permette de calculer le mode : on le trouve facilement à l'aide d'un tableau de pourcentages => c'est la valeur qui apparaît le plus souvent. On peut aussi le repérer à partir d'une représentation graphique. On parle de mode pour une variable discrète (pour toutes les variables qualitatives et pour les variables quantitatives discrètes). On parle de classe modale pour les variables quantitatives continues.

Nombre de personnes au sein des ménages en 1968 et en 1999

Nombre de personnes dans le ménage	proportion de ménages 1968	1999
1	21,0	30,9
2	26,6	31,1
3	18,6	16,1
4	15,0	13,7
5	9,3	5,6
6	4,8	1,6
7	2,4	0,6
8	1,2	0,2
9 et plus	1,2	0,2
Ensemble	100,0	100,0

Une précision : lorsqu'un seul score apparaît clairement comme le plus important, on dit que la variable est **unimodale**. Un diagramme en bandes verticales ne laissera voir

qu'une bande prononcée et sa valeur est donnée par l'axe des abscisses. Quand il y a, en revanche, deux modalités qui ont la même importance, on dira que la variable est **bimodale** (lorsque par exemple un diagramme présente une distribution à deux bosses, un peu à la manière d'un dos de chameau). Les deux bosses n'ont nullement besoin d'être à la même hauteur, il suffit qu'elles soient sensiblement égales entre elles et plus grandes que les barres des autres valeurs pour qu'on puisse dire que la variable est bimodale. C'est le cas de la structure des ménages en 1999 où le nombre de ménages composé d'une et de deux personnes est très proche (ce qui n'était pas le cas en 1968 où les ménages de deux personnes étaient la modalité la plus commune).

Parfois, on a une distribution plutôt plate, sans modalité concentrant une forte proportion de cas. Les modalités se répartissent de manière à peu près égale et aucune n'attire un nombre particulièrement élevé de cas. Dans ce cas, on parle d'une distribution faiblement modale. Auquel cas, le mode est de peu d'utilité.

A noter : le mode est la seule mesure qui convient à toutes les variables (peu importe qu'elle soit qualitative ou quantitative). C'est la seule mesure qui convienne aux variables qualitatives nominales.

N.B. : on repère le mode à partir des pourcentages ou à partir d'une représentation graphique mais on ne dit pas le mode est « … 437 (effectif) / 22 % (%) ». On indique la modalité de la variable correspondant au pourcentage le plus élevé (et non ce %).

II. La moyenne

La moyenne c'est la mesure de position la plus utilisée. On ne peut la calculer qu'avec des **variables quantitatives** (discrètes ou continues). On ne calculera jamais de moyenne pour les variables qualitatives, ça n'a pas de sens. On l'obtient en additionnant tous les scores et en divisant ensuite cette somme par le nombre total de scores. C'est relativement simple.

1. On additionne toutes les valeurs ;
2. On divise par la taille de la population.

Pour le calcul de la moyenne, on doit opérer une distinction entre les variables quantitatives discrètes et continues.

¤ Pour les variables quantitatives discrètes. Considérons le tableau suivant :

Variable	**Effectif**
x_1	f_1
x_2	f_2
…	…
x_n	f_n

La moyenne se calcule à partir de cette formule :

$$\bar{x} = \frac{\sum x_n f_n}{N} = \frac{x_1 \times f_1 + x_2 \times f_2 + ... + x_n \times f_n}{N} = \frac{\sum Valeur\ de\ la\ \mathrm{mod}alité \times Effectif\ de\ la\ valeur}{N}$$

où Σ (sigma) signifie additionner ou « somme » ;

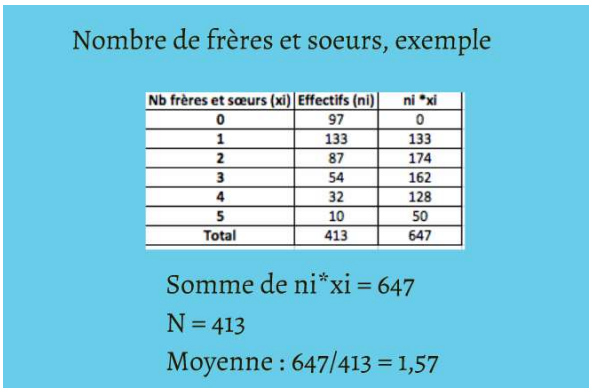
x_i représente chacune des modalités de la variable à tour de rôle ;
f_i représente chacun des effectifs de la variable x ;
N est la taille de la population.

La moyenne est donc la somme du produit des valeurs de la variable par leur effectif divisée par la taille de la population.

Si on reprend l'exemple de la taille des ménages et qu'on veut connaître leur taille moyenne :

$$\bar{x} = \frac{1 \times 155 \ + 2 \times 156 \ + \cdots + 9 \times 1}{502} = 2,41$$

Il y a quand même un problème à résoudre pour ce calcul. Quelle valeur doit-on attribuer à la catégorie « 9 personnes et plus » ? Il faut faire un choix « intelligent », en s'aidant de la distribution de la variable, ici je propose de prendre 9 personnes par ménage comme une valeur représentative de la classe. La taille moyenne des ménages français en 1999 est ainsi de 2,4 personnes par ménages.



¤ Pour les variables quantitatives continues, dans le cas de données groupées en classes, le point milieu de chaque classe correspond à la valeur qu'on prendra pour chaque classe. Ce point milieu sera utilisé pour représenter les données de sa classe. La moyenne se calcule donc en prenant le point milieu et la fréquence de chacune des classes.

La moyenne s'obtient à partir de cette formule :

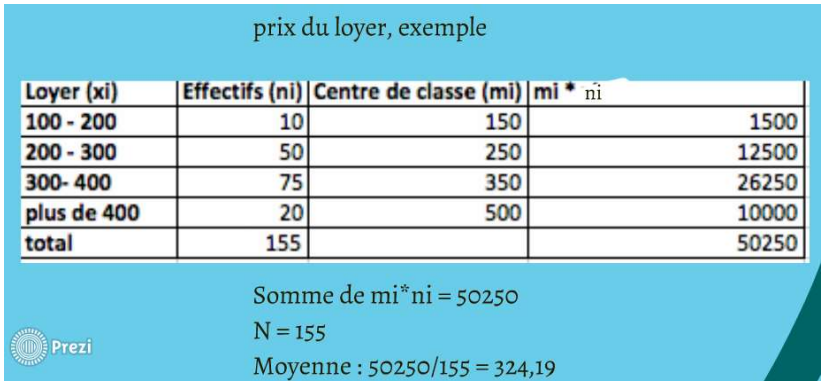
$$x = \frac{\sum m_i f_i}{N} = \frac{\sum (Point\ milieu\ de\ la\ classe) \times (Fréquence\ de\ la\ classe)}{N}$$

où Σ signifie « somme » ;

m_i représente chacun des points milieux des classe à tour de rôle ;

f_i représente l'effectif de la classe représente par mi ;

N est la taille de la population.



III. La médiane

La médiane est la valeur qui divise en deux groupes égaux une population. La médiane est le point en dessous duquel se trouve la moitié de la population et au-dessus duquel se trouve l'autre moitié de la population. Ainsi, pour une distribution de salaires, la médiane est le salaire

au-dessous duquel se situent 50 % des salaires. C'est aussi le salaire au-dessus duquel se situent 50 % des salaires. On l'abrège *Md*.

Elle suppose qu'il y ait une relation d'ordre croissant entre les modalités de la variable (sinon, ça n'a pas de sens). **Les modalités de la variable doivent être ordonnées**. On ne peut donc pas repérer la médiane pour toutes les variables : elle convient pour toutes les variables quantitatives et les variables qualitatives à échelles ordinales.

La médiane peut être un **meilleur indicateur que la moyenne** : de nombreuses grandeurs sont limitées vers le bas et non vers le haut. Par ex., le salaire horaire est limité vers le bas par le SMIC alors que certains salaires peuvent être très élevés. La moyenne est tirée vers le haut par les salaires élevés, même s'ils sont peu nombreux, et elle est pour cette raison souvent supérieure à la médiane. Celle-ci est de ce point de vue un indicateur plus fiable. Elle permet de se faire une idée de la forme de la distribution.

¤ **Le cas des variables (quantitatives) discrètes**

Dans le cas de variables discrètes, c'est simple. Deux manières de faire.

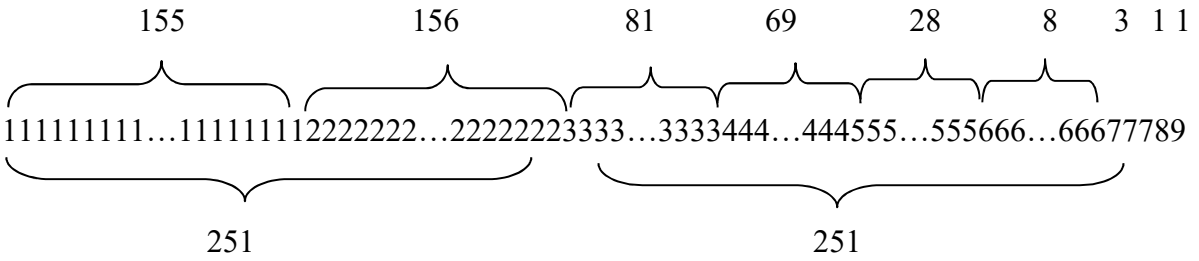
Déterminer la variable à partir d'un tableau statistique.

- 1. On dispose les modalités de la variable selon un ordre croissant.
- 2. On trouve la valeur de la variable qui partage la population en deux à partir des pourcentages cumulés croissants. Lorsque le seuil de 50 % vient d'être franchi, on regarde la valeur correspondante. La valeur médiane est celle vis-à-vis de laquelle on atteint le cumul de 50 % des données pour la première fois.

Nombre de personnes au sein des ménages en 1999

Nombre de personnes dans le ménage	Nombre de ménages		
	fréquences	pourcentages	pourcentages cumulés
1	155	30,9	30,9
2	156	31,1	62,0
3	81	16,1	78,1
4	69	13,7	91,8
5	28	5,6	97,4
6	8	1,6	99,0
7	3	0,6	99,6
8	1	0,2	99,8
9 et plus	1	0,2	100,0
Ensemble	502	100,0	

La médiane, c'est la valeur qui sépare les 502 données en deux blocs de 251 données chacun, c'est-à-dire en deux blocs contenant chacun 50 % des données. En observant les pourcentages cumulés, on voit que 50 % des ménages sont composés d'au plus 2 personnes et 50 % des ménages sont composés d'au moins 2 personnes. On choisit donc la valeur 2 comme médiane.



On ne peut pas dire qu'il y a 50 % des valeurs à gauche de 2, car il y a 155 données soit 30,9 % et, à sa droite, on ne peut pas le dire non plus car il y a 190 données, soit 37,85 %. Par contre, on peut dire qu'il y a au moins 50 % des données qui ont une valeur d'**au plus 2** et qu'il y a aussi au moins 50 % des données qui ont une valeur d'**au moins 2**. 2 est la **valeur médiane** de cette distribution.

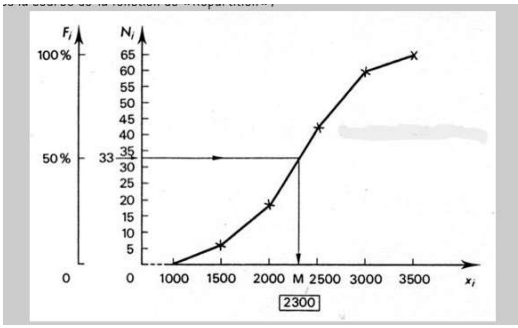
Cas particulier : si, avec les pourcentages cumulés, on tombe exactement sur 50 %. Ce qui signifie que la médiane se situe – si on imagine que cela se produisait dans l'exemple précédent – entre une donnée qui a 2 comme valeur et une autre donnée qui a 3 comme valeur. Dans ce cas, on peut dire : 50 % des ménages sont composés d'au plus deux personnes et 50 % des personnes sont composés d'au moins 3 personnes. Alors, quelle valeur choisira-t-on pour la médiane : 2 ou 3 ? Lorsqu'une telle situation se produit, la médiane, par convention, est le point milieu entre les deux valeurs => 2,5 personnes. Donc on interprétera la médiane en disant : au moins 50 % des ménages sont composés d'au plus 2,5 personnes.

Note : une telle situation peut se produire seulement avec un effectif pair et lorsque le pourcentage 50,0 % exactement est écrit dans la colonne des pourcentages cumulés.

Déterminer la variable à partir d'un graphique.

. Tracer une courbe des pourcentages cumulés croissants.

. Repérer la médiane : en traçant une parallèle à l'axe des abscisses au point d'ordonnées 50 %. De l'intersection de cette droite avec la courbe de répartition, on abaisse une perpendiculaire qui indique sur l'axe des abscisses la valeur médiane.



¤ **<u>Le cas des variables quantitatives continues</u>**

On peut, un peu comme dans la manière précédente, repérer la médiane à partir des pourcentages cumulés. Sauf que cette fois, on va repérer la **classe médiane** : celle vis-à-vis de laquelle on atteint le cumul de 50 % des données pour la première fois.

Il existe une méthode pour calculer de manière plus précise la médiane. On la calcule à partir des pourcentages. Elle se calcule selon cette formule :

$$Md = L + \left(\frac{50 - F}{f} \right) i$$

où L est la borne inférieure de la classe médiane ;
F est le pourcentage cumulé dans les classes précédent la classe médiane ;
f est le pourcentage des données dans la classe médiane ;
i est la largeur de la classe médiane (ou l'amplitude de la classe médiane).

Prenons l'exemple d'une distribution du revenu mensuel des ménages :

Revenus mensuels des ménages en 1998

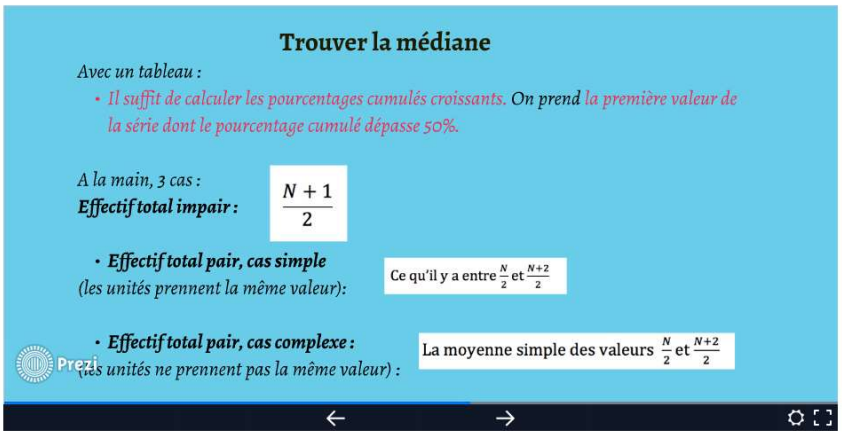
revenus mensuels	Nombre de ménages		
	fréquences	pourcentages	pourcentages cumulés
moins de 570	33	7,3	7,3
[570 ; 760[	34	7,6	14,9
[760 ; 1015[	48	10,7	25,6
[1015 ; 1270[	48	10,7	32,6
[1270 ; 1525[	55	12,2	48,4
[1525 ; 1905[	65	14,4	62,9
[1905 ; 2290[	49	10,9	73,8
[2290 ; 3050[	58	12,9	86,7
[3050 ; 3810[	32	7,1	93,8
3810 et plus	28	6,2	100,0
Ensemble	450	100,0	

La médiane se trouve dans l'intervalle compris entre 1525€ et 1905€ par mois. Son calcul exact devient :

$$Md = 1525 + \left(\frac{50 - 48{,}4}{14{,}4} \right) \times (1905 - 1525) = 1565{,}92$$

La médiane des revenus des ménages est donc de 1566 € par mois. C'est-à-dire que 50 % des ménages touchent au plus 1566 € par mois et 50 % des ménages touchent au moins 1566 € par moins.

Autre méthode pour déterminer la médiane :



IV. Les quantiles, les quartiles et les déciles

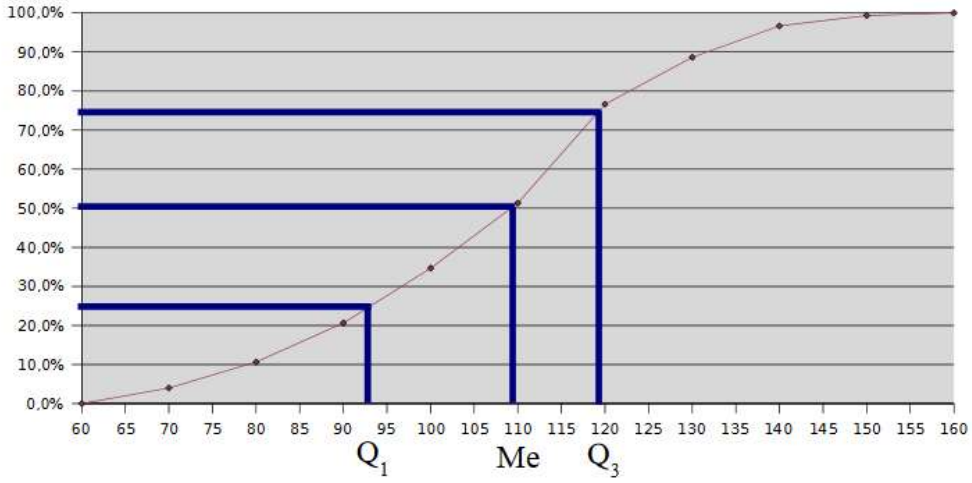
Déf quartiles : Si on ordonne une distribution de salaires, de revenus, de chiffre d'affaires..., les quartiles sont les valeurs qui partagent cette distribution en quatre parties égales. Ainsi, pour une distribution de salaires :
- le premier quartile (noté généralement Q1) est le salaire au-dessous duquel se situent 25 % des salaires ;
- le deuxième quartile est le salaire au-dessous duquel se situent 50 % des salaires ; c'est la médiane ;
- le troisième quartile (noté généralement Q3) est le salaire au-dessous duquel se situent 75 % des salaires.
Le premier quartile est, de manière équivalente, le salaire au-dessus duquel se situent 75 % des salaires ; le deuxième quartile est le salaire au-dessus duquel se situent 50 % des salaires, et le troisième quartile le salaire au-dessus duquel se situent 25 % des salaires.
http://www.inegalites.fr/spip.php?page=analyse&id_article=703&id_rubrique=110&id_mot=30&id_groupe=9

Un quantile correspond à une valeur où dessous de laquelle il y a un certain pourcentage de données ; dans ce sens la médiane est un quantile puisqu'elle donne la valeur au-dessous de laquelle il y a 50% des données. Les **quantiles** correspondent donc à des valeurs qui subdivisent le nombre de données en tranches égales, c'est-à-dire qu'entre deux valeurs le pourcentage de données est le même (par exemple, 25% de la population) sans que la distance entre les deux valeurs ne soient la même. On distingue surtout les **quartiles** et les **déciles**.

Les quartiles correspondent à des valeurs qui subdivisent le nombre de données placées en ordre croissant en tranches contenant chacune 25% des données. La notation utilisée pour les quartiles est Q_1, Q_2 (médiane) et Q_3.

Les **déciles** correspondent à des valeurs qui subdivisent le nombre de données placées en ordre croissant en tranches contenant chacune 10% des données. La notation utilisée pour les déciles est D_1, D_2,… D_9. Le cinquième décile correspond là encore à la médiane.

On utilise ce graphique pour déterminer le 1er quartile, la médiane et le 3ème quartile qui correspondent à des fréquances cumulées de 25%, 50% et 75%.



On lit ainsi :

- le 1er quartile est $Q_1 \approx 93$
- la médiane est Me ≈ 110
- le 3ème quartile est Q3 ≈ 119

Séance 8. Les indicateurs de dispersion

Préambule

Les mesures de tendance centrale ne sont pas suffisantes pour se faire une bonne idée de la forme d'une distribution. Prenons un exemple pour comprendre le problème. Supposons qu'Hugo, un jeune bachelier, envisage d'effectuer un voyage en bateau pendant ses vacances ; il a reçu deux propositions avec des conditions de voyage semblables. Il décide de faire son choix en se basant sur l'âge moyen de ses compagnons de voyage.

L'âge moyen des voyageurs du Moussaillon est de 22 ans ; celle des Quatre-vents est de 24 ans. Hugo choisit donc le Moussaillon. A-t-il fait le bon choix ?

Regardons ses potentiels compagnons de voyage et calculons la moyenne :

Le Moussaillon	Les Quatre-vents
Armand 37 ans	Pénélope 26 ans
Nadia 36 ans	Jérôme 25 ans
Germain 16 ans	Éléonore 24 ans
Florian 13 ans	Sylvain 24 ans
Amandine 8 ans	Camille 23 ans
	Basile 22 ans

Hugo passera donc ses vacances avec M. et Mme Festy et leurs trois enfants, et non avec une joyeuse bande d'amis. Ce qui a induit en erreur Hugo, et qui est la principale différence entre ces deux distributions, **c'est la dispersion des âges**.

A moyennes égales [trouver un exemple où les moyennes sont les mêmes ?], des séries statistiques peuvent donc présenter des **allures très différentes** : elles peuvent être très étalées (premier cas) ou très resserrées autour de la moyenne (2ème exemple). **Pour caractériser et résumer une distribution, il faut donc s'appuyer sur deux mesures** : d'une part, **les mesures de position qui informent sur son centre** (mesures de tendances centrales : mode, moyenne et médiane) et, d'autre part, **les mesures qui informent sur l'éparpillement des données autour de ce centre => ce qu'on appelle les indicateurs de dispersion**. Les indicateurs de dispersion résument en un seul nombre l'état de l'éparpillement ou l'éloignement des données autour du centre d'une distribution. **On mesure donc la dispersion autour d'une valeur et par convention, on utilise la moyenne comme point de référence pour décrire la dispersion** (on la choisit parce que contrairement au mode ou à la médiane, elle tient compte de toutes les valeurs).

Si les données divergent beaucoup, la distribution se présentera de façon étendue, le gros des données se situant loin de la donnée moyenne. Si les données ne divergent pas beaucoup, elles se masseront beaucoup plus près du centre et s'agglomèreront ensemble.

Je présenterai deux indicateurs de dispersion : la variance et l'écart-type. Ces mesures sont construites de sorte à tenir compte des **écarts** de toutes les valeurs par rapport à la moyenne.

I. La variance

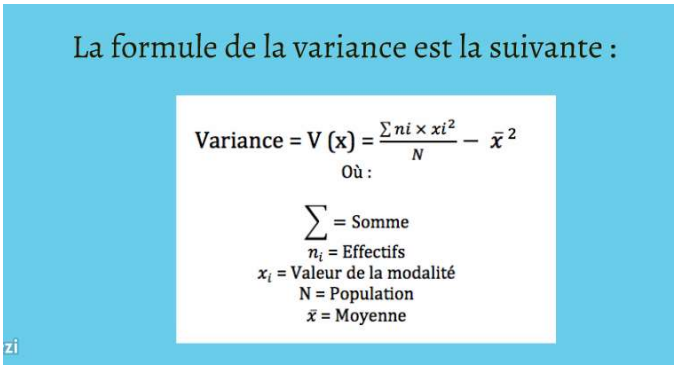
La variance d'une population est notée σ^2 (qui se prononce « sigma carré ») et elle se calcule selon cette formule : ajouter calcul variance avec le cas des var.quanti.continue où classes

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{N}$$

où x_i représente chaque modalité de la variable à tour de rôle ;
$\bar{x}$ représente la moyenne de la distribution ;
f_i représente chacun des effectifs de la variable x ;
N est la taille de la population.

Pour le dire plus simplement, il faut :

1. Calculer la moyenne ;
2. Soustraire la moyenne à chacune des valeurs de la variable (certaines de ces différences seront négatives et c'est normal) ;
3. Mettre au carré chacune de ces différences ;
4. Additionner toutes ces différences élevées au carré ;
5. Diviser cette somme par le nombre d'unités statistiques.



x_i représente chacune des modalités de la variable à tour de rôle ;

f_i (n_i)représente chacun des effectifs de la variable x ;
N est la taille de la population.

Dans le cas de classe, remplacez xi par mi (le centre de classe)

âge (xi)	effectifs (ni)	Centre de classe (mi)	ni *mi	ni * mi^2
8 à 10	50	9	450	4050
10 à 12	35	11	385	4235
12 à 15	110	13,5	1485	20047,5
15 à 18	120	16,5	1980	32670
18 à 20	45	19	855	16245
20 à 25	60	22,5	1350	30375
25 - 30	55	27,5	1512,5	41593,75
Total	475		8017,5	149216,25
Moyenne	16,87894737			
Variance	29,24060942			
écart types	5,407458684			

rezi

Quand les valeurs se distribuent de façon étendue autour de la moyenne, cette somme – c'est-à-dire la variance – sera grande ; quand, au contraire, les valeurs se distribuent de façon très resserrée par rapport à la moyenne, cette somme – c'est-à-dire la variance – sera faible.

Calculons la variance des âges des voyageurs sur les deux voiliers :

La variance de l'âge des voyageurs du Moussaillon sera :

¤ Formule 1.

$$\sigma^2 = \frac{[(37-22)^2 \times 1] + [(36-22)^2 \times 1] + [(16-22)^2 \times 1] + [(13-22)^2 \times 1] + [(8-22)^2 \times 1]}{5} = 146,8$$

¤ Formule 2.

V(x) = {[(37² * 1) + (36² * 1) + (16² * 1) + (13² * 1) + (8² * 1)] : 5} – 22² =

V(x) = [(1369 + 1296 + 256 + 169 + 64) / 5] – 22² = 3154 / 5 – 22² = 630,8 – 484 = 146,8

La variance de l'âge des voyageurs des Quatre-vents sera :

¤ Formule 1.

$$\sigma^2 = \frac{[(26-24)^2 \times 1] + [(25-24)^2 \times 1] + [(24-24)^2 \times 2] + [(23-22)^2 \times 1] + [(24-22)^2 \times 1]}{6} = 1,7$$

¤ Formule 2.

V(X) = {(26²*1)+(25²*1)+(24²*2)+(23²*1)+(22²*1) / 6} – 24²

V(X) = [(676 + 625 + 1152 + 529 + 484) / 6] - 576 = 3466 / 6 - 576 = 577,67 – 576 = 1,7

Dans la mesure où une forte variance témoigne d'une grande dispersion et où une faible variance témoigne d'une dispersion plus faible, avec cette seule mesure, Hugo se serait aperçu que les voyageurs du Moussaillon avaient des âges très éloignés de la moyenne contrairement à ceux des Quatre-vents.

II. L'écart-type

L'information est encore plus claire en calculant l'écart type. L'écart-type est la **racine carré de la variance**. L'écart-type d'une population se calcule donc de la manière suivante :

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2 n_i}{N}}$$

L'écart-type des âges des voyageurs du Moussaillon est de 12 alors qu'il n'est que d'1 pour les voyageurs des Quatre-vents. On dira que la plupart des voyageurs du Moussaillon ont un âge se situant à ± 12 ans de la moyenne, alors que la plupart des voyageurs des Quatre-vents ont un âge qui se situe à ± 1 an de la moyenne.
L'écart-type c'est l'indicateur de dispersion le plus utilisé mais on a besoin de la variance pour le calculer.

III. Interprétation de ces deux mesures

Comment interpréter ces deux mesures, la variance et l'écart type ?
Plus la variance et l'écart-type sont faibles, plus les valeurs de la variable sont centrées autour de la moyenne ; plus la variance et l'écart type sont grands, plus la distribution est dispersée. Ces deux indicateurs sont proches, mais il faut les distinguer. La variance a un inconvénient : comme elle met au carré les écarts à la moyenne, elle est difficilement interprétable. Il est plus facile de tenir compte de **l'écart type qui est calculé sur la même échelle que la moyenne**. Il exprime la dispersion des données dans la même unité que celle des données. L'écart-type peut prendre des valeurs allant de zéro à l'infini.

> Si l'écart-type est de zéro, cela signifie qu'il n'y a pas de dispersion, que toutes les données ont la même valeur.

Plus il y aura de données dont les valeurs s'éloignent de la valeur de la moyenne, plus la valeur de l'écart-type augmentera. Un écart-type élevé signifie une forte dispersion et une population **hétérogène**.

A contrario, un écart-type faible signifie que les valeurs sont concentrées autour de la moyenne et que la population regroupe des individus aux caractéristiques **homogènes**.

Il n'est pas facile de déterminer si les données sont plus ou moins dispersées en regardant la valeur de l'écart-type. Cette décision sera plus facile en comparant la valeur de l'écart-type d'une distribution avec celle d'une autre distribution.

Coefficient de variation ? Deux définitions :

https://fr.wikipedia.org/wiki/Coefficient_de_variation

http://www.insee.fr/fr/methodes/default.asp?page=definitions/coefficient-de-variation.htm