PREMIÈRE PARTIE
PRINCIPES ET MÉTHODES STATISTIQUES

Plan n° 1 — leçon

Introduction au rôle des statistiques en biologie humaine et dans la pratique des soins

INTRODUCTION

Sans toujours en être parfaitement conscient, on fait usage de nombreux concepts statistiques en santé publique et en médecine clinique chaque fois qu'il s'agit de prendre une décision sur l'un des points suivants: diagnostic clinique, prévision de l'impact probable d'interventions au niveau d'une collectivité ou déroulement d'une maladie chez un patient donné, choix de programmes d'intervention appropriés pour telle ou telle communauté ou d'un traitement pour tel ou tel patient, etc. Au laboratoire d'analyse, les statistiques sont inséparables de la pratique quotidienne. En outre, il est désormais essentiel de connaître les statistiques pour pouvoir comprendre et soumettre à une analyse critique les communications qui paraissent dans les revues médicales. Au total, la maîtrise des principes statistiques constitue une nécessité absolue pour la planification, l'exécution et l'analyse des études sur l'évaluation de la situation sanitaire et de ses tendances, ainsi que pour la conduite de travaux de recherche biomédicale, clinique ou en santé publique.

Objectif de la leçon

La présente leçon a pour objet de présenter aux étudiants le rôle des statistiques dans l'étude des populations humaines, de la biologie humaine et de la médecine et de les sensibiliser ainsi à la nécessité d'acquérir la connaissance des principes et des méthodes statistiques.

Connaissances à acquérir

A la fin de la leçon, l'étudiant devra savoir:

a) Discuter du rôle des statistiques dans la pratique des soins de santé et exposer les principales applications des méthodes statistiques dans le domaine des soins de santé, au sens large.

b) Indiquer au moyen d'exemples, en quoi les principes et les notions statistiques s'appliquent aux situations suivantes:

- prise en compte des variations des paramètres (par exemple physiologiques, chimiques) observés dans le domaine des soins de santé;
- diagnostic des troubles dont souffrent les patients et diagnostic des problèmes de santé au niveau communautaire;
- prédiction de l'effet probable à attendre de programmes d'interventions visant la maladie, soit au niveau communautaire, soit au niveau individuel;
- choix de modes de traitement appropriés en présence d'un patient donné;
- administration et planification de la santé publique;
- planification, exécution, analyse, interprétation et notification des travaux de recherche médicale.

Connaissances préalables requises

Les étudiants doivent:

- avoir une certaine expérience de la mesure et de l'utilisation de paramètres biologiques dont on sait qu'ils présentent une certaine dispersion;
- avoir une certaine connaissance des questions qui intéressent les systèmes médico-sanitaires et des grand objectifs des services de santé;
- savoir ce que signifient les mots diagnostic, pronostic et traitement;
- comprendre ce qu'on entend par science et par méthode scientifique.

Nouveaux termes et concepts

On trouvera ci-dessous la liste des nouveaux termes et concepts étudiés dans la présente leçon:

approches déterministe et probabilistique (stochastique); biostatistiques; collecte, réduction, synthèse, analyse et présentation des données; incertitude; méthodes statistiques; notions et principes statistiques; probabilité; sciences exactes et non exactes; statistiques; statistiques médicales; statistiques sanitaires; variation.

TENEUR DE LA LEÇON

Doivent être traitées les questions définies ci-dessous dans leurs grandes lignes.

Définitions des statistiques

Le terme «statistiques» est utilisé dans deux acceptions. Tout d'abord, il vise l'utilisation quotidienne:

- de données
- d'observations numériques
- de renseignements quantitatifs.

Exemples

1. Nombre d'agents de santé communautaires dûment formés dans les différentes subdivisions administratives du pays.
2. Poids de naissance des enfants.
3. Age (au dernier anniversaire) des patients vus un jour donné dans un dispensaire de soins ambulatoires.
4. Prévalence de la schistosomiase, rapportée à 1000 habitants, dans une circonscription administrative.
5. Taux de créatinine, en mg par litre, dans un échantillon d'urine des 24 heures.

Par ailleurs, le terme statistiques vise la *discipline* correspondante, laquelle englobe:

- la méthodologie statistique
- l'étude des méthodes scientifiques utilisées pour recueillir, traiter, réduire, présenter, analyser et interpréter les données et pour faire des déductions et tirer des conclusions à partir de données numériques.

Principales applications des méthodes statistiques

On peut faire trois usages principaux des méthodes statistiques:

a) Réunir des données de la meilleure façon possible

Cette application comprend les méthodes à utiliser pour:

- dessiner les fiches et imprimés pour la collecte des données
- organiser la collecte elle-même
- concevoir et exécuter un travail de recherche
- effectuer des enquêtes au niveau d'une population.

Exemples

1. Réunion de données sur les participants à un programme d'intervention visant une maladie.
2. Collecte systématique de données sur les naissances et les décès.
3. Collecte de données en vue de comparer les effets respectifs de l'administration, au troisième stade du travail, d'une association d'ergométrine et d'oxytocine, ou uniquement d'ergométrine.
4. Collecte de données sur les sujets atteints de tuberculose pulmonaire au sein d'une population déterminée.

b) Décrire les caractéristiques d'un groupe ou d'une situation

Cela suppose essentiellement:

- la réduction des données
- la synthèse des données
- la présentation des données.

c) Analyser les données et en tirer des conclusions

Il faut pour cela mettre en œuvre diverses techniques d'analyse puis appliquer les notions de calcul des probabilités pour en tirer des conclusions.

Application des concepts et principes statistiques dans la prestation des soins de santé

L'emploi des statistiques est essentiel dans la prestation des soins de santé, tant au niveau communautaire qu'au niveau individuel. La médecine s'intéresse à des sujets chez qui des paramètres tels que le poids, la taille, la tension artérielle, le taux de cholestérol, le taux d'immunoglobuline, la glycémie, etc., présentent une certaine dispersion. L'état de bonne santé pour tel ou tel

paramètre correspond à une valeur variable selon les individus. Alors qu'il n'existe pas de sujet, ni de groupes de sujets, qui soient rigoureusement identiques, les décisions concernant les patients ou la communauté doivent être fondées sur des observations faites auprès d'autres patients ou d'autres communautés semblables sur le plan biologique et social. Force est de reconnaître qu'en présence de telles différences, le bien-fondé des décisions n'est pas indiscutable: ces décisions sont entachées d'une certaine incertitude. C'est en cela que consiste la **nature probabiliste de la médecine.**

Il est donc nécessaire d'être parfaitement familiarisé avec les techniques qui conviennent pour faire face à ces différences et aléas.

De plus, l'application des statistiques est utile pour acquérir l'esprit critique indispensable:

- pour réfléchir aux problèmes médicaux de façon scientifique, logique et critique
- pour apprécier correctement les observations sur lesquelles il est possible de fonder la décision
- pour être conscient des risques possibles que comportent toute décision médicale
- pour repérer les décisions et les conclusions qui manquent d'un fondement scientifique et logique.

Les principes et concepts statistiques sont appliqués dans divers domaines de la médecine. En voici quelques exemples.

a) Prise en compte des variations

Un paramètre (ou facteur, ou mesure) est variable lorsqu'il n'a pas la même valeur chez tous les sujets ou à tous les instants, chez un sujet donné. La quasi-totalité des paramètres observés dans la prestation des soins de santé, qu'ils soient physiologiques, biochimiques ou immunologiques, sont variables.

Exemples. Age, poids, taille, tension artérielle, taux de cholestérol, taux de bilirubine, d'albumine, d'immunoglobuline, nombre de plaquettes, glycémie.

On est donc confronté à des problèmes chaque fois qu'on essaie de trouver une valeur qui résume les observations faites au sujet d'un paramètre dans un groupe de patients ou au niveau d'une communauté, c'est-à-dire de choisir pour le paramètre la valeur idéale, normale, moyenne, etc., ou encore de comparer deux groupes de patients, ou deux communautés, du point de vue d'un paramètre donné. C'est uniquement lors-

qu'on a clairement défini ces problèmes qu'on peur choisir des méthodes statistiques appropriées pour les résoudre.

b) *Diagnostic des troubles des patients et de l'état de santé d'une collectivité*

Par diagnostic, on entend la démarche qui permet de définir l'état de santé d'un sujet, ou d'un groupe de sujets, et repérer les facteurs qui en sont la cause. Bien souvent, c'est par des méthodes reposant implicitement sur les techniques statistiques qu'on a défini les diverses entités morbides en se fondant sur le regroupement de signes, symptômes et valeurs de certains paramètres biochimiques.

Lorsqu'on décide que la santé d'un individu ou d'une collectivité doit les faire rattacher à l'une de ces entités, il existe toujours une certaine incertitude. Il se peut que les signes et symptômes observés ne correspondent pas exactement à ceux dont la liste sert de définition à l'entité. Inversement, la même série de signes et symptômes peut relever de plusieurs entités morbides.

Quand un agent de santé choisit l'entité morbide qui a le plus de chance d'être correcte, il fait, inconsciemment, un raisonnement statistique. Il existe des méthodes statistiques explicites qui permettent de classer les diverses entités envisageables, en indiquant pour chacune d'elle la probabilité qu'elle a de correspondre à la réalité.

c) *Prédiction du résultat probable d'un problème d'intervention au niveau d'une maladie, dans une communauté ou chez un individu*

Par pronostic, on entend l'évaluation ou la prédiction des résultats probables d'un programme d'intervention dans une communauté ou chez un patient, compte tenu des signes d'appel et des circonstances. On s'appuie sur les observations précédemment faites lors de programmes d'intervention similaires, de sorte que la démarche est en principe essentiellement de nature statistique.

Il est indispensable d'avoir consigné les observations faites lors de l'examen initial et au cours du traitement ainsi que l'issue de la maladie au niveau communautaire ou chez les patients précédemment vus par le clinicien. L'analyse des données ainsi consignées permet de voir en détail quelle a été l'issue de la maladie chez les différents sujets. En s'appuyant sur cette analyse, on peut établir le résultat probable d'un nouveau programme d'intervention ou d'un nouveau traitement individuel.

d) *Choix d'une intervention appropriée pour un patient ou pour une communauté*

On s'appuie pour cela:

— sur les observations antérieures faites chez les patients ou au sein de collectivités semblables ayant bénéficié de cette intervention.

— sur les rapports parus dans la littérature au sujet d'essais ou d'expériences cliniques visant à apprécier l'efficacité relative de différents médicaments ou autres thérapies.

— sur l'évaluation objective des observations antérieurement faites par l'agent de santé.

Si l'on veut que les conclusions soient valables, il faut que la conception, l'exécution de l'analyse de l'expérimentation et des interventions médicales soient conduites selon des principes et des méthodes statistiques rationnelles. Sans cette condition, une intervention risque d'être inefficace, voire nocive à l'insu du médecin.

e) *Administration et planification en matière de santé ou de santé publique*
Dans ce domaine, la principale application consiste à utiliser les données sur la pathologie en cause dans la population en vue de formuler un diagnostic au niveau communautaire. Il faut pour cela connaître:

— des caractéristiques telles que l'effectif et la structure d'âge de la population;

— le profil sanitaire de la population, en termes de distribution des facteurs de maladies ou de risques;

— l'influence des facteurs écologiques;

— le maniement des biostatistiques (données sur les naissances et les décès).

En outre, l'administration et la planification sanitaires font appel à des données sur la distribution des soins de santé à tous les niveaux de l'analyse (besoins, disponibilité, utilisation, etc.).

Tous les éléments ci-dessus sont mesurés à l'aide de taux ou d'autres indices statistiques; les gestionnaires de la santé doivent savoir les obtenir, les calculer, les utiliser et les interpréter. Vu que les agents de santé occupent la place principale dans l'établissement et dans l'utilisation de ces statistiques, ils ont une responsabilité primordiale dans leur exactitude.

f) *Planification, exécution, analyse et notification des recherches médicales: lecture et compréhension des communications dans ce domaine*
Tous les travaux dans le domaine médical, qu'ils prennent la forme d'analyses ou d'enquêtes descriptives, supposent la collecte, l'analyse et l'interprétation correctes de données numériques pertinentes. La validité de ces

études est subordonnée à l'application de principes statistiques rationnels à tous les stades.

La littérature médicale abonde en comptes rendus d'études; l'examen des publications actuelles montre à l'évidence qu'on fait un usage croissant des concepts et méthodes statistiques. S'ils veulent rester à la pointe du progrès dans leur profession, les agents de santé doivent être capables de lire et comprendre ces rapports et d'en faire l'analyse critique.

PLAN DE LA LEÇON

L'exposé peut être articulé comme suit. Tout au long de la leçon, on utilisera abondamment des exemples pris dans la littérature médicale actuelle ainsi que dans d'autres publications (y compris la presse quotidienne, le cas échéant) pour illustrer la nature des informations importantes dans le domaine de la santé.

a) Comme introduction générale à la statistique à l'intention des étudiants en médecine, examiner les objectifs généraux et précis du cours dans son ensemble, en soulignant qu'il ne s'agit pas de préparer des statisticiens sanitaires mais des agents de santé comprenant la nature des connaissances qu'ils ont à utiliser et des décisions qu'ils auront à prendre. Présenter une vue d'ensemble du cours, de sa structure, de son organisation et de ses méthodes et de sa répartition dans le temps.

b) Expliquer la signification des termes «statistiques» et «méthodes statistiques» et les termes apparentés, en donnant des exemples d'application à la médecine, en particulier: données numériques; informations quantitatives; méthodes convenables pour leur collecte, traitement, analyse, présentation, interprétation et diffusion; implications de la «médecine scientifique» et de la «méthode scientifique».

c) Expliquer le rôle essentiel des statistiques dans le domaine de la santé (par exemple dans la connaissance médicale, la pratique médicale, etc.) en examinant et en analysant des exemples sur la nature et les modalités des décisions prises par des agents de santé dans l'accomplissement de leurs tâches (à savoir dans le diagnostic, le pronostic et le traitement), et des décisions touchant à l'administration, à la planification et à l'évaluation en matière sanitaire.

d) Examiner les problèmes que soulèvent les variations et incertitudes pour l'étude des causes des maladies, des facteurs étiologiques ou des facteurs de risque; l'évaluation de la réponse thérapeutique; l'établissement des

valeurs «normales», «habituelles» et «idéales» des divers paramètres et, par conséquent, la nature des méthodes à appliquer à leur étude.

e) Montrer que, pour les raisons examinées ou illustrées antérieurement, la généralisation de l'emploi des méthodes statistiques est évidente dans les communications qui paraissent de nos jours dans les revues médicales. C'est dans cette littérature que les futurs médecins puiseront l'information voulue pour mettre à jour leurs connaissances théoriques et pratiques. Il leur faut donc avoir une connaissance suffisante des principes et méthodes statistiques de base pour pouvoir évaluer la validité et la fiabilité de l'information recueillie dans les rapports médicaux et les articles scientifiques. De plus, il faut qu'ils soient familiarisés avec la terminologie technique des méthodes statistiques dont ils trouveront mention dans la littérature.

Plan n° 2 — leçon

Collecte et organisation des données et échelles de mesure

INTRODUCTION

Pour conduire de façon systématique et permanente la formulation, la planification, la programmation, la budgétisation et la mise en œuvre des politiques de santé et pour procéder à l'intégration d'ensemble des différents programmes dans le système global de santé, il faut pouvoir disposer d'informations. Les méthodes utilisées pour réunir et analyser les données dépendent des usagers potentiels et de la nature de l'information qui leur est a priori nécessaire. La qualité de cette information dépend toujours du mode de collecte des données et de la nature de l'«instrument» (méthode, formules, appareil de mesure, etc.) employé pour la collecte. La quantité d'information qu'on peut tirer des données réunies dépend de la nature de l'échelle de mesure utilisée.

Objectif de la leçon

L'objectif de la présente leçon est de donner une idée aux étudiants de la nature et des types de données nécessaires pour soutenir le processus gestionnaire pour le développement de la santé, du mode de collecte des données et des diverses échelles utilisables pour leur mesure.

Connaissances à acquérir

A la fin de la leçon, l'étudiant devra savoir:

a) Distinguer, parmi les systèmes utilisables pour collecter des données relatives à la santé, les systèmes habituels et les systèmes *ad hoc*.

b) Décrire les méthodes utilisables pour la collecte des données sanitaires.

c) Expliquer quels sont les différents types d'«instruments» de mesure.

d) Expliquer les concepts de fiabilité et de validité, dans le contexte de la métrologie, et en discuter les implications pour l'exploitation des données sanitaires.

e) Faire la différence entre les quatre principales échelles de mesure et indiquer leur application respective pour la collecte de données sanitaires.

f) Distinguer les données quantitatives et les données catégorielles.

Connaissances préalables requises

Signification des termes «statistiques», «méthodes statistiques» et des termes connexes et application dans le domaine de la médecine, comme on l'a vu dans le premier plan (leçon).

Nouveaux termes et concepts

On trouvera ci-dessous la liste des nouveaux termes et concepts étudiés dans la présente leçon:

attribut; description qualitative; description quantitative; données catégorielles; données quantitatives; échelle continue; échelle discontinue; échelles de mesure; fiabilité; mesure objective; mesure subjective; sensibilité; sources *ad hoc* de données; sources de données de routine; spécificité; statistiques descriptives; validité; variable.

TENEUR DE LA LEÇON

Systèmes de routine et systèmes *ad hoc* pour la collecte des données

Un système *de routine* employé pour la collecte des données comprend en général un mécanisme (une méthode d'enregistrement) qui assure la collecte des données à mesure qu'elles sont disponibles.

Exemples

1. Registres d'état civil pour la collecte de données sur les naissances, les décès, les mariages et les divorces.
2. Système de notification des maladies pour la collecte de données sur le choléra, la fièvre jaune, la coqueluche, etc.
3. Système de notification des cas de cancer (registre du cancer).
4. Systèmes d'enregistrement dans les établissements de soins pour la collecte de données sur les patients qui sont vus dans les différents services.

La collecte *ad hoc* de données prend généralement la forme d'une enquête destinée à rassembler des renseignements dont on ne dispose pas de façon

régulière. Cette enquête peut comporter l'étude analytique spéciale ou l'approfondissement de certains aspects de données qui sont réunies de façon systématique. Les données rassemblées peuvent être destinées à des fins administratives ou de recherche.

Exemples

1. Une enquête nationale sur les personnels de santé.
2. Une enquête visant à estimer la proportion d'enfants malnutris au sein d'une population déterminée.
3. Une étude visant à établir si l'utilisation de contraceptifs hormonaux influe sur l'état nutritionnel.
4. Une étude sur la pratique de l'allaitement maternel chez les femmes qui ont déclaré une naissance l'année précédente.

Méthodes de collecte des données

a) Système de routine

Pour réunir des données de façon régulière, on procède en général comme suit (mais pas nécessairement dans l'ordre indiqué):

- Fixation de règles et règlements instituant le système et lui donnant un support juridique, spécialement lorsqu'il s'agit d'un système au niveau national. Ces règles et réglements sont promulgués par l'organe législatif ou l'autorité compétente.
- Installation matérielle des bureaux, recrutement du personnel et diffusion de l'information appropriée dans le grand public.
- Choix des éléments d'information à recueillir.
- Conception des formules, fiches et registres qui serviront à l'enregistrement des données.
- Formation du personnel.
- Définition de la procédure d'enregistrement: identité des personnes chargées de fournir l'information, moment où cette information doit être enregistrée, etc.
- Définition et conception des reçus d'enregistrement, autrement dit des pièces délivrées à la personne qui vient faire enregistrer une naissance, un décès, etc., pour témoigner qu'elle s'est acquittée de son obligation. On peut donner comme exemples les fiches d'enregistrement à l'hôpital, les certificats de déclaration d'une naissance ou d'un décès, etc.

b) *Système* ad hoc:

L'organisation d'une collecte exceptionnelle de données comporte les étapes suivantes:

- Définition ou énoncé des objectifs de la collecte, avec indication du type de renseignements à réunir et de leur mode d'utilisation.
- Définition de la population sur laquelle on a besoin de ces données (population de référence ou population cible).
- Décision sur l'ampleur de la collecte, auprès de toutes les unités de la population de référence ou de certaines d'entre elles seulement.
- Choix du nombre de «répondants» ou «enquêtés» (c'est-à-dire de sujets auprès desquels on va réunir les données) à inclure dans l'étude.
- Choix de ces enquêtés.
- Elaboration des instruments (imprimés, etc.) à utiliser pour l'enregistrement des données.
- Choix et formation du personnel chargé de recueillir les données.
- Collecte des données: identification des unités et des enquêtés choisis, rédaction des imprimés, etc.

Instruments de mesure

Il existe trois principaux types d'instruments de mesure:

- *Appareils:* la mesure est effectuée à l'aide d'un dispositif purement mécanique.

  Exemples. Balance, thermomètre, spectrophotomètre, sphygmomanomètre.
- *Opérateurs:* la mesure est effectuée par des personnes (pratiquement) sans utilisation d'appareil.

  Exemples. Auscultation du cœur, cotation d'une splénomégalie, anamnèse.
- *Combinaison* d'opérateurs et d'appareils.

  Exemples. Interprétation d'une radiographie, lecture d'un étalement sanguin.

Fiabilité et validité

Deux caractéristiques sont souhaitables pour les instruments de mesure, la fiabilité et la validité.

Fiabilité

La fiabilité concerne les performances intrinsèques d'un instrument. Est fiable (fidèle) un instrument qui donne des résultats uniformes lorsqu'on refait plusieurs fois de suite la mesure sur la même unité et dans des conditions similaires. La fidélité dépend principalement des facteurs suivants:

- Variation inhérente à l'instrument lui-même (fidélité).

  Exemples. Fluctuation du zéro dans une balance, absence de stabilité des réactifs utilisés pour construire un instrument mécanique.

- Fluctuations au niveau de l'objet de la mesure.

  Exemples. Réponses différentes fournies par un même patient aux questions qu'on lui pose lors de l'anamnèse, influence sur la réponse du degré de compréhension des questions.

- Erreur d'observation: un même observateur peut obtenir des résultats différents lorsqu'il recommence la mesure sur le même objet.

  Exemples. Seconde mesure de la tension artérielle, nouvelle détermination de l'âge (quand on ignore la date de naissance), nouvelle numération des microfilaires sur une lame colorée.

- «Equation personnelle»: différences entre les valeurs obtenues par plusieurs observateurs.

  Exemples. Mesures de la tension artérielle, interprétation de radiographies, lecture d'étalements sanguins.

Validité

Une mesure est valable si elle reflète l'état qu'elle est censée mesurer.

Exemples

1. La fièvre peut être un indicateur non valable (insuffisant) du paludisme dans les régions où cette parasitose a un faible degré de transmission.
2. Les réponses fournies lors d'interrogatoires directs (oralement) risquent, dans certaines sociétés, de ne pas révéler la pratique réelle locale en matière d'avortement.
3. L'absence d'enfant n'est pas toujours un indice valable de stérilité.

La validité comporte deux aspects importants: la sensibilité et la spécificité.

La *sensibilité* d'une épreuve, d'une méthode ou d'un instrument de mesure est égale, de façon générale, au quotient de la variation de la mesure observée à la variation correspondante de la valeur de la grandeur ou du paramètre qui fait l'objet de la mesure. Plus ce quotient est élevé, plus la sensibilité est importante. Par exemple, lorsqu'il s'agit de mesurer une concentration et qu'une petite variation de la concentration se traduit par une variation importante du résultat fourni par une épreuve, on dit que cette épreuve est sensible. Ainsi entendue, la sensibilité ne désigne *pas* la plus petite quantité ou valeur mise en évidence par une méthode donnée (valeur que l'on désigne en toute rigueur sous le nom de «seuil de détection»). En épidémiologie, la sensibilité est définie par *la proportion des vrais positifs correctement repérés* par une épreuve; elle se calcule par la formule $a/(a+c)$ dans laquelle a=nombre de vrais positifs correctement identifiés et c=nombre de faux négatifs donnés par épreuve.

Par définition la *spécificité* exprime la mesure dans laquelle une épreuve, une méthode ou un instrument de mesure réagit à la présence d'une variable donnée et reste «insensible» à la présence de toutes les autres variables. En épidémiologie, la *spécificité* est égale à *la proportion des vrais négatifs correctement identifiés* par une épreuve; elle se calcule au moyen de la formule $d/(b+d)$, dans laquelle d=nombre de vrais négatifs correctement identifiés et b=nombre de faux positifs.

Dans les épreuves de dépistage, la validité comporte d'autres aspects: la valeur prédictive positive ou négative. La première est égale à la probabilité pour qu'un résultat positif fourni par l'épreuve indique un résultat authentiquement positif. La seconde est égale à la probabilité pour qu'un résultat négatif fourni par l'épreuve indique un résultat authentiquement négatif. Par exemple, dans le cas d'une maladie, il s'agit de la probabilité d'avoir la maladie si l'on présente le signe ou d'être indemne si on ne le présente pas. Ces divers paramètres sont plus faciles à comprendre à l'aide du Tableau 2.1.

Tableau 2.1. Valeurs prédictives, positive et négative

		Situation réelle +	Situation réelle −	Total
Résultats de l'épreuve	+	a	b	$a+b$
	−	c	d	$c+d$
	Total	$a+c$	$b+d$	

Sensibilité $= a/(a+c)$
Spécificité $= d/(b+d)$
Valeur prédictive positive $= a/(a+b)$
Valeur prédictive négative $= d/(c+d)$

Variables et attributs

Une variable ou *grandeur mesurable* correspond à une grandeur qui prend diverses valeurs numériques, soit d'un sujet à l'autre, soit, pour un même sujet, d'un instant à l'autre.

Exemples. La taille exprimée en mètre ou le poids exprimé en kilogramme.

Un *attribut* descriptif représente, pour une grandeur donnée, une catégorie à laquelle le sujet appartient ou n'appartient pas ou une qualité que le sujet possède ou ne possède pas.

Exemples. L'accès à une certaine forme de soins de santé, la maladie, l'hospitalisation, le groupe sanguin A.

Certaines grandeurs ne peuvent être considérées que d'une façon seulement, tandis que les autres peuvent l'être des deux façons. Par exemple, le poids corporel peut être étudié en tant que variable (poids en kg) ou en tant qu'attribut (surcharge pondérale/absence de surcharge). Le choix entre ces deux optiques dépend de la raison de la mesure, des conditions à remplir en matière d'objectivité, de fiabilité et de validité et des propriétés des différentes échelles de mesure utilisables. Ces divers aspects sont étudiés plus loin.

Variables continues et variables discrètes

Une variable continue est une variable qui peut a priori prendre un nombre infini de valeurs à l'intérieur d'un intervalle donné. Elle peut prendre des valeurs entières ou fractionnaires et peut être mesurée avec une précision variable selon le degré de raffinement de la méthode utilisée.

Exemples. Hauteur (en mètres): 1,83, 1,74; poids (en kg): 48,7, 90.

Une variable discrète ne peut prendre qu'un nombre fini de valeurs à l'intérieur d'un intervalle donné. Ces valeurs sont en général (mais pas toujours) des nombres entiers.

Exemples. Nombre d'enfants d'une famille; nombre de ménages d'une communauté; nombre de leucocytes; nombre de lits dans une salle d'hôpital.

Echelles de mesure

Les quatre principales échelles utilisées pour mesurer des données sont l'échelle nominale, l'échelle ordinale, l'échelle graduée et l'échelle graduée avec zéro.

Une *échelle nominale ou échelle de classification* consiste dans l'emploi de noms ou de qualificatifs divers pour distinguer une «mesure» d'une autre. La mesure sur ce type d'échelle n'implique aucune notion de grandeur.

Exemples.

1. L'issue de la maladie chez un sujet donné a deux mesures possibles: la survie ou le décès.
2. L'engagement national vis-à-vis des soins de santé primaires peut être jugé comme étant réel ou inexistant.
3. Les malades mentaux peuvent être classés d'après la nature de leur trouble: psychose, névrose, dépression maniaque ou schizophrénie.

Une *échelle ordinale* présente les caractéristiques de l'échelle nominale décrite ci-dessus mais implique l'existence d'une relation d'ordre dans l'ensemble des mesures.

Exemples.

1. L'absence d'une alimentation correcte pour les mères allaitantes et les enfants dans une région frappée par la sécheresse peut être classée en plusieurs catégories selon qu'elle est critique, grave, modérée ou légère.
2. La situation sociale des patients peut être mesurée par la classe à laquelle ils appartiennent: élite, classe moyenne ou classe inférieure.

Une *échelle graduée* se caractérise par une unité numérique de mesure qui fait que la différence entre deux mesures s'exprime explicitement sous forme d'un multiple (ou sous-multiple) d'un intervalle unitaire, séparant deux points de l'échelle. Une caractéristique particulière de cette échelle est que l'unité de mesure et le zéro (origine ou point de départ) sont arbitraires, c'est-à-dire fixés de façon conventionnelle.

Exemples. En général, la température corporelle se mesure sur une échelle graduée, en utilisant par exemple comme unité le degré Celsius (°C).

L'échelle ordinale peut prendre l'aspect d'une échelle graduée si l'on attribue un score (une valeur numérique) aux diverses catégories de cette échelle; elle n'en conserve pas moins les propriétés d'une échelle ordinale.

L'échelle graduée avec zéro présente toutes les caractéristiques de l'échelle graduée mais elle présente en plus un zéro vrai ou zéro absolu de sorte que le rapport entre deux valeurs de l'échelle constitue une mesure significative de l'importance relative de ces deux mesures.

Exemples. La taille en m ou le poids en kg.

Certaines opérations arithmétiques peuvent être valablement effectuées sur chacune de ces échelles.

Echelle nominale. Sur cette échelle, une opération possible est celle d'«équivalence»: par exemple, une «femme» équivaut à une autre «femme». Les mesures équivalentes peuvent être regroupées à l'intérieur d'une catégorie particulière et être comptées. On peut calculer la proportion des mesures appartenant à une catégorie donnée par rapport au nombre total de mesures.

Echelle ordinale. Sur cette échelle, une mesure peut être égale à une autre (équivalente) ou au contraire supérieure ou inférieure. La différence entre deux mesures n'est pas explicite et des différences entre mesures voisines ne sont pas équivalentes. Là encore, les mesures équivalentes peuvent être regroupées au sein d'une catégorie donnée et comptées et l'on peut calculer la proportion des mesures qui tombent dans les diverses catégories.

Echelle graduée. Toutes les opérations possibles sur l'échelle ordinale le restent sur l'échelle graduée, mais on peut en outre additionner ou soustraire les mesures ou les diviser ou multiplier par une constante en obtenant des résultats interprétables. La comparaison d'intervalles pris sur cette échelle a un sens et ne dépend pas de l'unité de mesure ni du système d'attribution de scores.

Echelle graduée avec zéro. Toutes les opérations arithmétiques sont possibles et le rapport de deux mesures quelconques a un sens et ne dépend pas de l'unité choisie.

Données quantitatives et données catégorielles

Les données peuvent être réparties en deux grandes catégories selon l'efficacité de l'échelle de mesure:

Les *données catégorielles* sont des mesures dans lesquelles la notion de grandeur est absente ou implicite. Les variables correspondantes se mesurent soit sur une échelle nominale, soit sur une échelle ordinale. Ces données sont également désignées sous le nom de données qualitatives ou attributives.

Les *données quantitatives* ont une grandeur. Elles se mesurent sur une échelle graduée, avec ou sans zéro.

PLAN DE LA LEÇON

L'exposé peut être articulé comme suit. Dans la mesure du possible, on tirera des exemples de la littérature pour illustrer les divers points traités.

a) Expliquer la signification et l'importance des «statistiques descriptives» et leur place dans l'information et les connaissances médicales.

b) Examiner les différentes sources de données et les systèmes et méthodes utilisés pour leur collecte. Etudier leurs caractéristiques (utilité, disponibilité, qualité et coût).

c) Faire la différence entre collecte de routine (systématique) et collecte *ad hoc*. Montrer en quoi diffèrent les collectes de données qui ont essentiellement une finalité statistique et celles qui sont destinées à l'administration ou à la gestion de la santé. En suivant la hiérarchie administrative du système de santé, montrer à l'aide d'exemples quelles sont les données réunies aux différents niveaux. Indiquer quelles sont les données transmises à l'échelon supérieur, sous quelle forme et à quelle fin (voir l'exemple du polycopié 2.1, p. 25). Montrer qu'il est indispensable que l'information circule dans les deux sens, en amont et en aval à l'intérieur de la hiérarchie administrative et du système de soins de santé. Insister sur le fait qu'un mécanisme efficace de retransmission vers l'amont constitue un aspect important de tout système d'appui informationnel fonctionnant à point nommé, et de façon efficace et rentable.

d) La description ou la mesure est à la base même de toutes les statistiques descriptives. Faire la différence entre variables et attributs, description quantitative et description qualitative, critères de mesure objectifs et subjectifs, données quantitatives et catégorielles.

e) Exposer les quatre types d'échelle de mesure et expliquer leurs propriétés eu égard à la quantité d'information transmise, à la fiabilité et à la validité.

EXERCICES EN CLASSE

Demander aux participants:

a) d'énumérer les problèmes de santé publique qu'ils ont rencontrés ainsi que leurs caractéristiques (épidémiologiques, biologiques). Pour chaque caractéristique, ils devront indiquer:

- l'échelle de mesure;
- la nature, discrète ou continue, de cette caractéristique;
- le type d'instrument utilisé pour la collecte de données sur ce sujet.

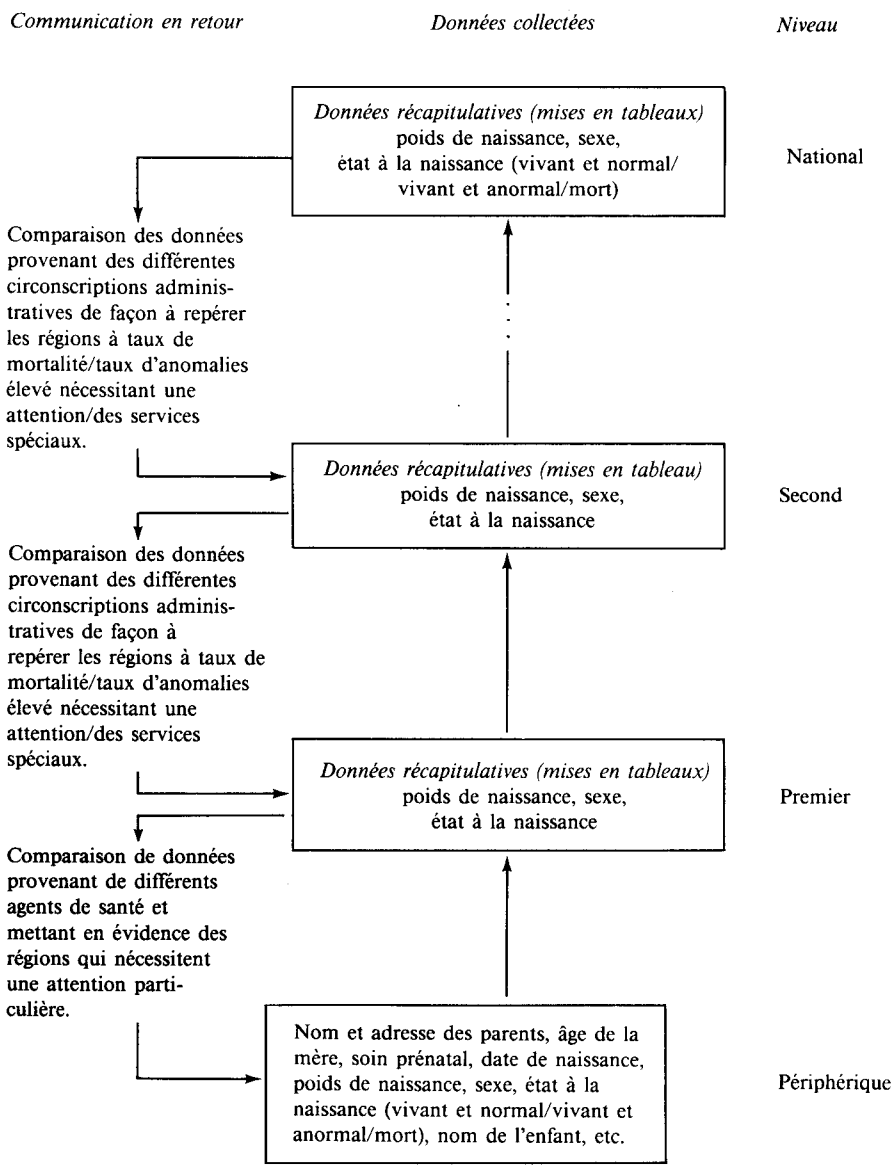
b) d'énumérer cinq types d'information réunie dans le pays et relative à la santé/médecine communautaire:

— par le canal d'un système ordinaire;

— à l'aide d'enquêtes *ad hoc*.

c) de décrire trois situations dans laquelle on effectue des mesures, chacune avec l'un des types d'instruments décrits plus haut. Préciser quels sont les facteurs risquant de compromettre la fiabilité de l'instrument et expliquer comment on peut en tenir compte.

POLYCOPIÉ 2.1 **Exemple de circulation de l'information sanitaire au sein du système de santé dans le cas d'un programme de santé maternelle et infantile**



Plan n° 3 — leçon

Présentation des données

INTRODUCTION

Dans la plupart des cas, il n'est pas facile de distinguer du premier coup d'œil les renseignements utiles dans une masse de données brutes. Il faut organiser les données réunies de façon à condenser l'information contenue pour que les types de variations se dégagent clairement. On ne peut choisir une méthodologie précise en vue de l'analyse que lorsqu'on a compris les caractéristiques de la série de données.

Objectif de la leçon

L'objectif de la présente leçon est de faire comprendre aux étudiants pourquoi il est indispensable de condenser et d'organiser la présentation des données et de leur donner une idée des techniques utilisables pour ce faire.

Connaissances à acquérir

A la fin de la leçon, l'étudiant devra pouvoir:

a) Exposer les circonstances dans lesquelles la condensation et la présentation de données sanitaires est indispensable.

b) Reconnaître les avantages et les inconvénients respectifs de la présentation des données sous forme de tableaux et de diagrammes.

c) Expliquer les applications et le mode d'établissement des présentations suivantes:

- série ordonnée;
- tableau de distribution (effectifs, fréquences et fréquences cumulées);
- tableau à plusieurs entrées;
- histogramme;
- polygone de fréquence;
- courbe des fréquences cumulées (sigmoïde);

— diagramme en bâtons;
— diagramme à secteurs.

d) Mettre en tableaux une série de données à l'aide d'une méthode appropriée.

e) Appliquer une méthode appropriée pour présenter sous forme de diagrammes des données qui le sont sous forme de tableaux.

f) Donner au moins trois exemples de mauvaise utilisation de la méthode des diagrammes pour la présentation des données.

Connaissances préalables requises

Leçons 1 et 2, spécialement en ce qui concerne les types de données et les types d'échelles de mesure.

Nouveaux termes et concepts

Les nouveaux termes et concepts dont la liste suit doivent être abordés dans la présente leçon (voir les définitions dans le polycopié 3.1):

classe; diagramme en bâtons; diagramme à secteurs; effectif; effectif cumulé; fréquence; fréquence cumulée; histogramme; intervalle de classe; limites de la classe; limites réelles de la classe; polygone des fréquences; présentation sous forme de diagrammes; présentation sous forme de tableaux; série ordonnée; sigmoïde; tableaux à plusieurs entrées; tableau des fréquences.

TENEUR DE LA LEÇON

Le référer, pour la teneur de la leçon, aux polycopies 3.1, 3.2 et 3.3.

PLAN DE LA LEÇON

Tout au long de la leçon, on choisira des exemples pour illustrer les différents modes de présentation des données. Certains des exemples peuvent s'inspirer des indications fournies dans le polycopié 3.2. La leçon peut être articulée comme suit.

a) Expliquer la notion de distribution d'une variable. Donner des exemples pris dans la littérature médicale pour illustrer les différentes «formes» de distribution et leurs implications du point de vue causal ou étiologique (courbe en cloche ou en chapeau de gendarme, courbe en J, courbe en L, distribution unimodale, distribution bimodale, courbe symétrique, courbe asymétrique, etc.).

b) Expliquer l'utilisation d'une série ordonnée pour faire apparaître une distribution à l'intérieur d'un petit ensemble d'observations.

c) Expliquer l'utilisation d'un tableau de fréquences pour mettre en évidence la distribution et montrer comment on établit ce type de tableau. Parler, en particulier, des points suivants:

- Nombre de classes, qui doivent être suffisamment élevées pour faire apparaître la distribution.
- Rapport entre l'intervalle de classe (étendue de la classe) et le nombre de classes, compte tenu des valeurs extrêmes.
- Utilisation de classes égales ou inégales selon le type de distribution.
- Utilisation de classes ouvertes à une extrémité pour y faire entrer les valeurs extrêmes de la distribution et faire en sorte que la classification soit complète aux deux extrémités de la distribution (même si les classes ne contiennent aucune observation).
- Définition correcte des intervalles de classe de façon que les classes ne se chevauchent pas mais qu'aucune valeur ne soit omise.

d) Expliquer l'utilisation d'une distribution de fréquences (par exemple une distribution de pourcentage) pour comparer deux distributions ou plus.

e) Exposer l'utilisation des tableaux à plusieurs entrées pour obtenir la distribution de fréquences d'une variable en fonction de sous-ensembles d'une autre variable (par exemple âge et sexe).

f) Décrire l'utilisation d'une distribution des fréquences cumulées et montrer comment on peut l'obtenir à partir d'un tableau de fréquences de base. On insistera particulièrement sur le fait que la définition des classes est modifiée dans un tableau des fréquences cumulées.

g) Expliquer (ou rappeler le point *a*) ci dessus) que les distributions apparaissent mieux et sont plus faciles à comparer quand les données sont présentées sous forme de diagrammes et plutôt qu'en tableaux, en se référant spécialement aux diagrammes ci-dessous (dont on indiquera le mode de construction à partir de données présentées sous forme d'un tableau de fréquences):

- *Histogramme.* La fréquence est représentée correctement par l'*aire* des rectangles jointifs, mais des précautions s'imposent lorsqu'on reporte sur le diagramme les fréquences d'un tableau qui comporte différents intervalles de classes ou certaines classes ouvertes à une extrémité. Expliquer comment il faut graduer et désigner l'axe horizontal (échelle de mesure) et l'axe vertical (effectifs ou fréquences, selon le cas).

— *Polygone des fréquences.* Ce mode de présentation permet de superposer deux distributions de fréquences et de les comparer facilement sur un graphique (alors que cette comparaison n'est pas toujours claire avec des histogrammes).

— *Diagramme des fréquences cumulées.* Parfois désignée sous le nom de sigmoïde (ou ogive de Galton), cette présentation est obtenue à partir d'un tableau des fréquences cumulées. On veillera particulièrement à situer les points du diagramme de façon que l'impression visuelle corresponde à la réalité.

— *Diagramme en bâtons.* Les barres peuvent être tracées verticalement ou horizontalement. Elles ont toutes la même largeur de sorte que les fréquences sont simplement proportionnelles à la longueur.

— *Diagramme en secteurs.* Il est utilisé pour comparer la fréquence relative de différents groupes. La fréquence est représentée par la surface, laquelle est proportionnelle à l'angle au sommet du secteur.

h) Expliquer et montrer de quelle façon un diagramme peut donner une représentation faussée des faits ou être utilisée à mauvais escient, par exemple lorsqu'on essaie de faire apparaître trop d'éléments sur un seul diagramme, lorsqu'on omet le zéro de l'échelle, lorsqu'on représente de façon incorrecte des intervalles inégaux. Voir des exemples au polycopié 3.2.

EXERCICES EN CLASSE

Demander aux participants:

a) De noter, pour chaque autre participant au cours, son sexe, son âge, son poids, sa taille, puis d'organiser ces données sous forme:

— de distributions de fréquences

— de tableaux à plusieurs entrées

— de diagrammes de fréquences.

b) De comparer les avantages et les inconvénients respectifs du diagramme en bâtons et du diagramme à secteurs.

POLYCOPIÉ 3.1. **Définitions des nouveaux termes et concepts**

Classe	Intervalle subdivisant l'intervalle global de variation du paramètre étudié: chacun des intervalles 3,0-3,3, 3,4-3,7,..., 5,0-5,3 constitue une classe.
Classification ou regroupement	Subdivision de l'intervalle de variation du paramètre étudié en une série de classes ou de groupes.
Diagramme en bâtons	Mode de présentation des effectifs d'une série de classes nominales par des bâtons de longueur proportionnelle à ces effectifs.
Diagramme à secteurs	Représentation de classes nominales sous forme de secteurs circulaires d'aire proportionnelle à l'effectif de la classe.
Distribution des effectifs (ou fréquences) cumulés	Distribution du nombre d'observations cumulées jusqu'à l'extrémité de la classe particulière considérée. Pour l'obtenir, on peut procéder de proche en proche.
Effectif de classe	Nombre d'observations faites dans chaque classe.
Fréquence de classe	Rapport du nombre d'observations appartenant à la classe au nombre total d'observations.
Histogramme	Diagramme présentant la distribution des effectifs ou des fréquences d'une variable quantitative par des rectangles d'aire proportionnelle à ces effectifs ou fréquences.
Limites de la classe	Valeur de la variable qui délimite chaque classe: par exemple, 3,0 et 3,3 sont respectivement la borne inférieure et la borne supérieure de la classe 3,0-3,3.
Limites réelles de la classe	Valeurs extrêmes rattachées à une classe donnée. Quand la mesure est effectuée à un dixième près, les limites de la classe 3,0-3,3 sont 2,95 et 3,34.
Polygone des fréquences	Diagramme représentant la distribution des fréquences d'une variable quantitative par une ligne brisée qui joint les extrémités d'un diagramme en bâtons obtenu en traçant, au niveau de la valeur centrale de chaque classe, un segment (bâton) de longueur proportionnelle à l'effectif ou fréquence de la classe.

Série ordonnée	Simple réaménagement des observations individuelles par ordre croissant ou décroissant.
Sigmoïde	Forme graphique la plus fréquente d'une distribution des fréquences cumulées.
Tableau ou distribution des effectifs ou fréquences	Tableau indiquant le nombre d'observations (en valeur absolue ou relative) qui correspondent à des caractéristiques particulières dans une série de données.
Tableau à plusieurs entrées	Tableau de fréquences contenant au moins deux variables classées sans des entrées différentes.

POLYCOPIÉ 3.2 **Exemples de présentation des données**

Exemple 1

Lors d'une enquête portant sur 50 femmes, on a observé les taux suivants de sérumalbumine (en grammes par litre de sang):

42,	41,	42,	44,	44,	36,	38,	41,	42,	44
42,	39,	49,	40,	45,	32,	34,	43,	37,	39
41,	39,	48,	42,	43,	33,	43,	35,	32,	34
39,	35,	43,	44,	47,	40,	39,	42,	41,	46
37,	49,	41,	39,	43,	42,	47,	48,	51,	52

En utilisant 5 intervalles égaux, on peut présenter ces données sous forme d'une distribution de fréquence comme suit:

Tableau 3.1. Distribution de fréquence des taux de sérumalbumine

Taux de sérumalbumine (g/litre)	Nombre d'observations
30-33	3
34-37	7
38-41	14
42-45	17
46-49	7
50-53	2
Total	50

Ce mode de présentation fait apparaître la distribution des taux de sérumalbumine chez les 50 personnes étudiées: on voit que le taux varie de 30 à 53 et qu'il se situe dans un nombre appréciable de cas entre 38 et 45 g/litre.

Exemple 2

Dans une étude (théorique) sur la relation entre la cheilite commissurale et la profession, on a rangé les sujets étudiés en cadres ou membres de professions libérales (C), ouvriers qualifiés (Q) et ouvriers non qualifiés (NQ). Au total, 88 sujets étaient atteints de cheilite commissurale (+) et 100 en étaient indemnes (−). On peut présenter la liste des sujets étudiés en fonction de leurs caractéristiques professionnelles et morbides, sous la forme du tableau 3.2.

Tableau 3.2. Mise en tableau des données sur la cheilite commissurale

Sujet	Cheilite commissurale	Profession
1	−	C
2	+	NQ
3	+	NQ
4	−	Q
.	.	.
.	.	.
.	.	.
186	+	Q
187	−	NQ
188	−	C

Ces données peuvent être présentées sous forme d'un tableau à deux entrées, l'une correspondant à la cheilite commissurale (présente ou absente) et l'autre à la profession (cadres, ouvriers qualifiés, ouvriers non qualifiés).

Tableau 3.3. Distribution des 188 sujets d'après leur catégorie professionnelle et la présence ou l'absence de cheilite commissurale

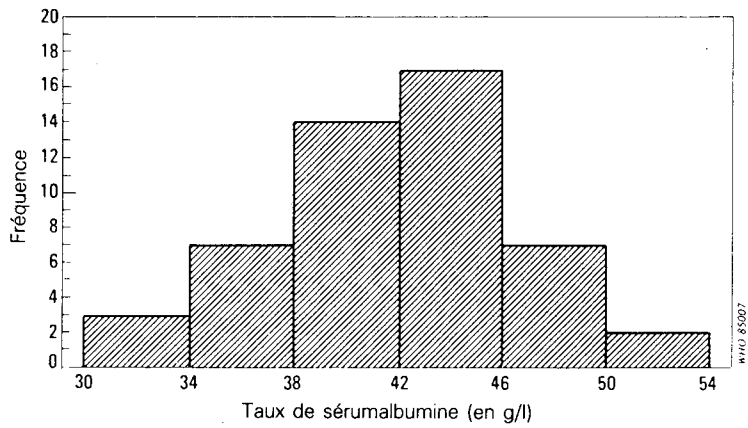
Cheilite commissurale	Catégorie professionnelle			Total
	Cadres	Ouvriers qualifiés	Ouvriers non qualifiés	
Présente	5	13	70	88
Absente	20	30	50	100
Total	25	43	120	188
Pourcentage de sujets malades	20,0	30,2	58,3	46,8

Le Tableau 3.3 fait apparaître la fréquence de cette affection dans les diverses catégories professionnelles.

Exemple 3

Les observations concernant le taux de sérumalbumine (Exemple 1) peuvent également être présentées graphiquement, sous forme d'un histogramme (Fig. 3.1).

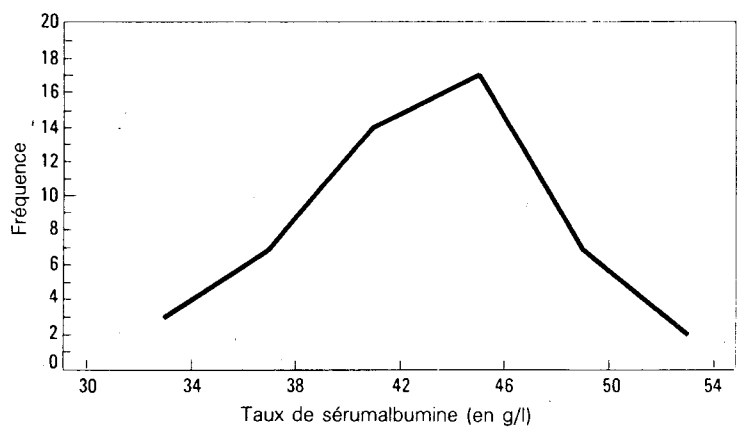
Fig. 3.1. Histogramme des taux de sérumalbumine observés dans l'exemple 1.



Exemple 4

Les taux de sérumalbumine observés à l'exemple 1 peuvent encore être présentés graphiquement sous forme d'un polygone de fréquence (Fig. 3.2).

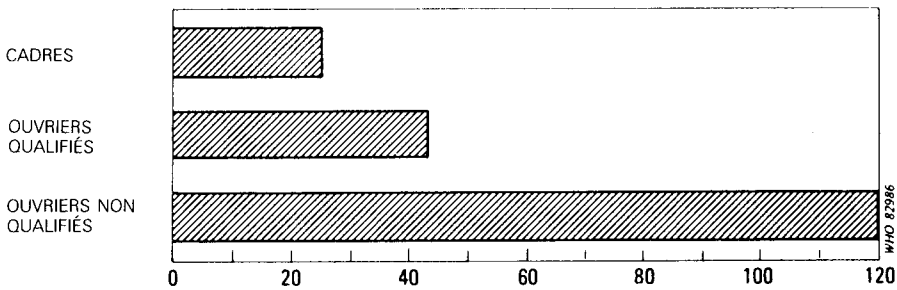
Fig. 3.2. Polygone de fréquences des taux de sérumalbumine observés dans l'exemple 1.



Exemple 5

La distribution professionnelle de 188 sujets de l'exemple 2 peut être présentée sous forme d'un diagramme en bâtons (Fig. 3.3).

Fig. 3.3. Diagramme en bâtons des catégories professionnelles observées à l'exemple 2.



Exemple 6

On peut utiliser un diagramme à secteurs (Fig. 3.4) pour représenter la distribution des catégories professionnelles de l'exemple 2. Ici, les fréquences doivent être converties en angles qui leur soient proportionnels.

Catégorie professionnelle	Fréquence	Angle au centre
Cadres	25	A
Ouvriers qualifiés	43	B
Ouvriers non qualifiés	120	C
Total	188	360°

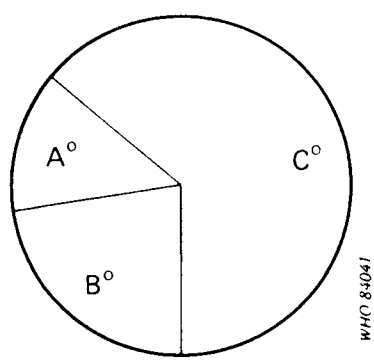
Fig. 3.4. Présentation des catégories professionnelles de l'exemple 2 sous forme d'un diagramme à secteurs.

$$\frac{A}{360^\circ} = \frac{25}{188} \text{ etc.}$$

D'où: $A = \dfrac{25 \times 360^\circ}{188} = 48^\circ,$

$B = \dfrac{43 \times 360^\circ}{188} = 82^\circ,$

$C = \dfrac{120 \times 360^\circ}{188} = 230^\circ.$

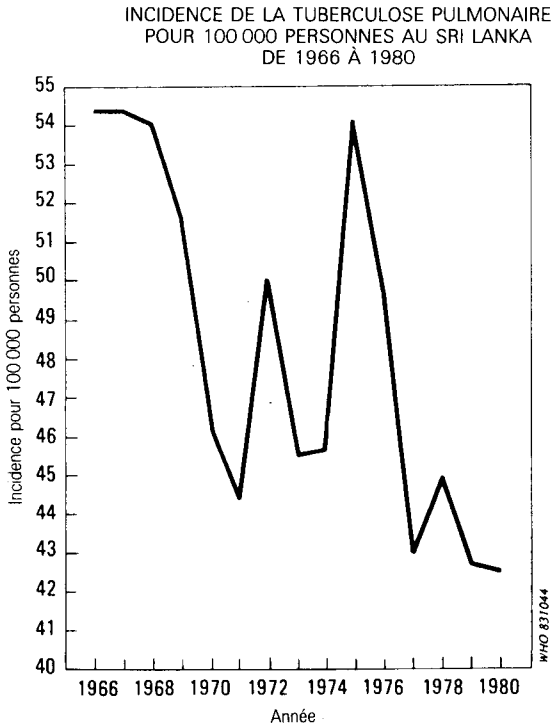


Exemple 7

Les figures 3.5, 3.6, 3.7 et 3.8 donnent des exemples de présentation graphiques fausses ou risquant d'induire en erreur.

Fig. 3.5. Erreur dans le choix de l'échelle

a)



b)

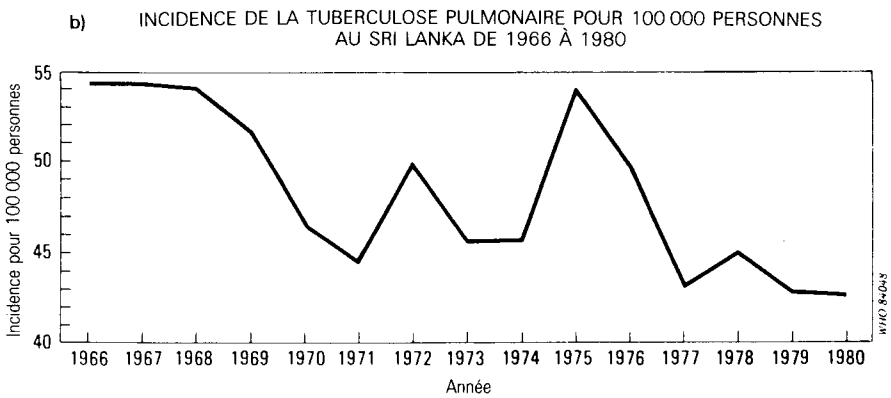


Fig. 3.6. Absence de point origine (zéro) sur l'axe vertical

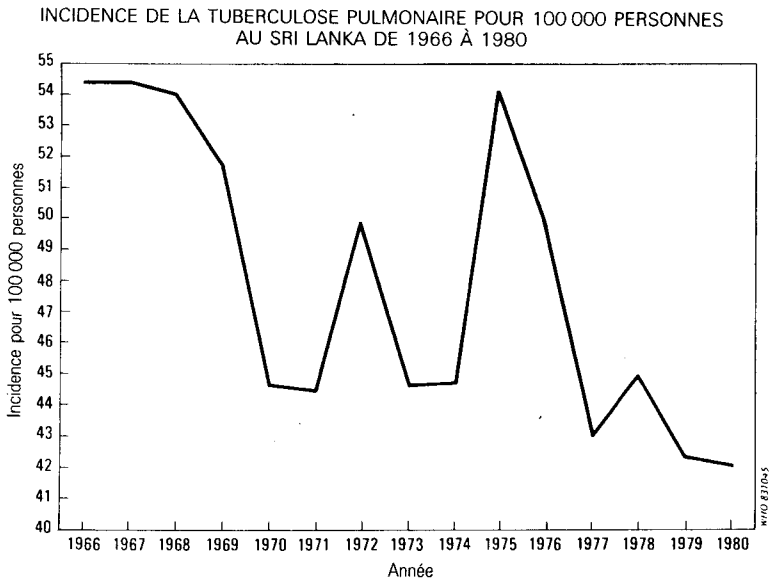


Fig. 3.7. Choix d'intervalles égaux sur l'axe horizontal ne correspondant pas à des variations égales du paramètre considéré

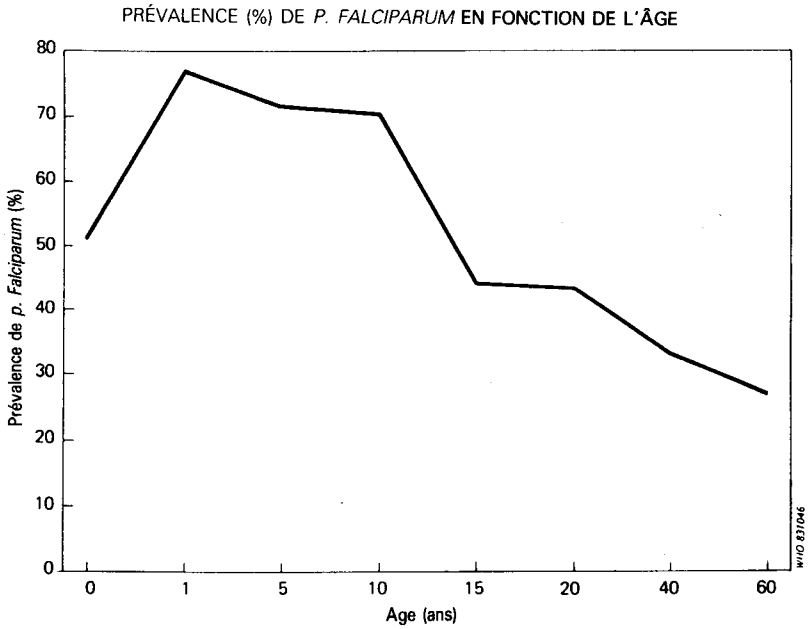
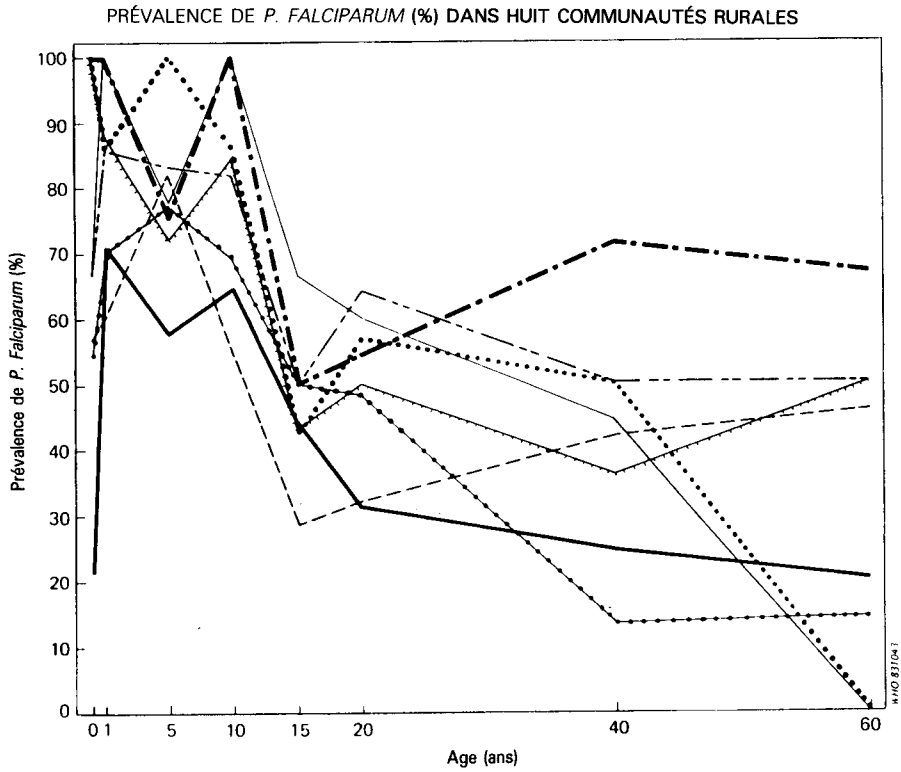


Fig. 3.8. Graphique surchargé



POLYCOPIÉ 3.3 **Choix de la solution la mieux adaptée entre une présentation tabulaire et une présentation graphique**

On trouvera ci-dessous les principales situations qui conviennent le mieux à l'usage de tel ou tel mode de présentation.

Méthodes tabulaires

Caractéristiques des données	*Méthode tabulaire*
Petit ensemble de données, par exemple moins de 20	Série ordonnée
Observations individuelles nombreuses mais portant sur une seule variable	Tableau de fréquences
Observations individuelles portant sur deux variables ou plus	Tableau à plusieurs entrées

Modes de représentation graphique adaptés à des données tabulaires

Données tabulaires	*Mode de représentation graphique*
Tableau de fréquences, variable quantitative, une série de données	Histogramme ou polygone de fréquences
Tableau de fréquences, variable quantitative, deux séries de données	Polygone de fréquences
Tableau de fréquences, données catégorielles	Diagramme en bâtons ou à secteurs

Plan n° 4 — leçon

Caractéristiques de tendance centrale ou de position

INTRODUCTION

La variabilité est une caractéristique propre à toutes les mesures. S'agissant de décrire une série de mesures ou de comparer plusieurs séries, il est impossible de décrire, ou de comparer, valablement les valeurs observées sur tous les membres de la «population» étudiée. Il est indispensable de définir les indices synthétiques appropriés. L'un de ces indices permet de repérer la position «centrale» (par exemple la moyenne des valeurs observées) ou la valeur la plus caractéristique de toutes les observations faites. Il s'agit de caractéristiques de tendance centrale ou de position.

Objectif de la leçon

La présente leçon a pour objectif de définir, étudier, les indices de valeur centrale (moyenne, médiane, mode et centiles), leur utilisation, leur interprétation et leurs limites.

Connaissances à acquérir

A la fin de la leçon, l'étudiant devra savoir:

a) Expliquer pourquoi on a besoin d'indices synthétiques en médecine.

b)† Calculer la moyenne, la médiane et le mode d'une série de données (groupées ou non groupées).

c) Expliquer les utilisations et les limites de la moyenne, de la médiane et du mode ainsi que leurs avantages et inconvénients respectifs en tant qu'indice synthétique, récapitulant des données sanitaires.

d) Expliquer l'utilisation des centiles pour récapituler des données sanitaires.

e) Choisir un paramètre approprié de valeur centrale pour une application donnée.

f) Faire la différence, dans le cas de données sanitaires, entre valeurs «moyenne», «normale» et «idéale».

Connaissances préalables requises

Toutes les questions exposées dans les leçons précédentes.

Nouveaux termes et concepts

Les nouveaux termes et concepts ci-après devront être enseignés lors de cette leçon. On distribuera aux étudiants un polycopié regroupant les définitions et les explications de façon qu'ils puissent s'y reporter pendant le cours et après. Devront être traités les points suivants:

caractéristiques de tendance centrale ou de position; centiles; distribution bimodale; distribution multimodale; indices synthétiques; médiane; mode; moyenne arithmétique; moyenne pondérée.

Ces nouveaux termes et concepts sont définis dans le polycopié 4.1, joint en annexe à la présente leçon.

TENEUR DE LA LEÇON

Le professeur devra articuler son cours en s'aidant des éléments contenus dans les polycopiés proposés et dans les exemples ci-dessous.

Il faudra utiliser tout au long du cours des exemples comme ceux qui suivent ou qui figurent dans le polycopié proposé sous 4.2.

Exemple 1. Moyenne

Lors d'une enquête sur la dimension des ménages dans un village, on a observé les effectifs suivants:

5, 3, 9, 7, 1, 3, 6, 8, 2, 6, 6.

Pour calculer la moyenne, on commence par additionner les chiffres ci-dessus:

$$5+3+9+\ldots+6=56.$$

On divise alors le total ainsi obtenu (56) par le nombre d'observations (11).

On obtient ainsi la moyenne: $56/11=5{,}1$ personnes.

Exemple 2. Médiane

Pour calculer la taille médiane des ménages de l'Exemple 1, on commence par classer les valeurs par ordre croissant (on obtient ainsi une série ordonnée):

1, 2, 3, 3, 5, 6, 6, 6, 7, 8, 9.

La valeur qui divise cette distribution en deux moitiés est la sixième de la série ordonnée, qui correspond à un effectif de six personnes. C'est donc la valeur médiane pour cette distribution. Quand le nombre d'observations est pair, la médiane se calcule en faisant la moyenne des *deux* valeurs centrales de la série.

Exemple 3. Mode

Si l'on reprend la série ordonnée des tailles des ménages de l'Exemple 2, à savoir:

1, 2, 3, 3, 5, 6, 6, 6, 7, 8, 9,

on constate que le chiffre qui revient le plus souvent est 6. Pour cette série, le mode est égal à 6.

Exemple 4. Moyenne pondérée

Dans différents villages, les enfants qui ne fréquentent pas encore l'école ont en moyenne l'âge suivant (en mois).

Tableau 4.1. Age moyen des enfants d'âge préscolaire dans différents villages

Village	Nombre d'enfant	Age moyen (mois)
1	54	58,6
2	52	59,5
3	49	61,2
4	48	62,5
5	48	64,5

Pour calculer l'âge moyen pondéré des enfants qui ne vont pas encore à l'école, on multiplie l'âge observé dans chaque village par l'effectif correspondant (coefficient de pondération) et l'on divise le total obtenu par l'effectif global:

$58{,}6 \times 54 + 59{,}5 \times 52 + 61{,}2 \times 49 + 62{,}5 \times 48 + 64{,}5 \times 48 = 15\,353{,}2$ mois,

$54 + 52 + 49 + 48 + 48 = 251$.

L'âge moyen pondéré est donc égal à $15\,353{,}2/251 = 61{,}2$ mois.

PLAN DE LA LEÇON

L'exposé peut être articulé comme suit:

a) Faire un bref rappel sur les aspects intéressants de la présentation des données, en particulier les distributions de fréquences, le regroupement des données et les intervalles de classe et les caractéristiques des distributions de fréquences.

b) Expliquer les nouveaux termes et concepts en s'aidant du polycopié 4.1. Tout au long de la leçon, on devra illustrer par des exemples simples les utilisations des paramètres de valeur centrale dans le domaine médical. Chaque fois qu'on cite une valeur «normale» en médecine, il s'agit d'un indice de la variable considérée: c'est ainsi qu'on parle de température «normale», de poids «normal» pour l'âge, etc. Etudier les implications des qualificatifs *idéal, habituel, optimal, type et usuel* par rapport à la moyenne, la médiane et le mode. Par exemple, la valeur du poids indiqué pour un groupe de personnes (qui peuvent souffrir de sous-nutrition) peut être la moyenne pour ce groupe mais ne pas correspondre à un chiffre normal. De même, une valeur normale n'est pas forcément idéale s'agissant d'un fonctionnement optimal et de la protection contre la maladie.

c) Expliquer les limites des indices, en prenant des exemples dans le domaine médical. On insistera sur le fait que la médiane et le mode sont peu sensibles à certains aspects d'une distribution de données mais qu'ils s'agit de paramètres utiles pour décrire une distribution asymétrique. Il faudra également expliquer l'influence des valeurs extrêmes sur la moyenne.

d) Lors de l'étude du calcul des indices, insister sur les principes de base plutôt que sur la mémorisation des formules. On soulignera les différences de méthodes selon qu'il s'agit de données groupées ou non groupées. On expliquera que, dans le second cas, les indices obtenus représentent directement la série de données correspondantes tandis que les valeurs calculées après regroupement des données ne sont qu'approchées.

Il faut expliquer clairement comment on détermine la valeur centrale d'une classe et quelles sont les hypothèses à faire pour les intervalles ouverts à une extrémité. On examinera l'influence de la précision des données sur la valeur estimative de la moyenne et de la médiane (calculées à partir de données groupées).

EXERCICES EN CLASSE

Tableau 4.2. Fréquentation quotidienne d'un centre de santé en novembre 1980

Date	Nombre de consultations	Date	Nombre de consultations	Date	Nombre de consultations
1	60	11	80	21	72
		12	71	22	89
3	77	13	108		
4	65	14	68	24	75
5	90	15	102	25	80
6	80			26	88
7	105	17	85	27	90
8	80	18	79	28	77
		19	97	29	99
10	85	20	87		

Demander aux étudiants:

a) De calculer la moyenne, la médiane et le mode du nombre de visites quotidiennes.

b) D'utiliser les valeurs ainsi obtenues pour commenter la distribution.

POLYCOPIÉ 4.1 **Définitions des nouveaux termes et concepts**

Caractéristique de tendance centrale ou de position	Indice synthétique décrivant la valeur «centrale» ou la valeur la plus caractéristique d'une série de mesures.
Centiles (ou percentiles)	Valeurs qui, dans une série d'observations classées par ordre croissant, répartissent la distribution en 100 parties égales (la médiane correspondant au 50^{e} centile).
Distribution multimodale	Distribution de données qui comporte plusieurs modes. (Quand il y a deux modes, on parle de distribution *bimodale.*)
Indice synthétique	Valeur donnant une vue globale d'une série d'observations.
Médiane	Valeur qui divise une distribution en deux moitiés égales; quand les valeurs observées sont classées par ordre de grandeur, la médiane correspond à la valeur centrale.
Mode	Valeur qui est observée le plus fréquemment dans une série.
Moyenne arithmétique	Valeur obtenue en faisant la somme de toutes les valeurs d'une série d'observations et en divisant le total par le nombre d'observations.
Moyenne pondérée	Valeur moyenne obtenue en pondérant (c'est-à-dire en multipliant par un coefficient de pondération) les diverses valeurs de la série; le coefficient choisi est très souvent la fréquence de l'observation.

POLYCOPIÉ 4.2 **Exemple de calcul de la moyenne et de la médiane pour des données groupées**

Tableau 4.3. Tension artérielle systolique de 240 sujets de sexe masculin

Tension artérielle systolique en mmHg (intervalle de classe)	Fréquence (f)	Valeur centrale de la classe (x)	Produit (fx)	Fréquence cumulée
Moins de 100	4	95[a]	380	4
100–	16	105	1 680	20
110–	18	115	2 070	38
120–	40	125	5 000	78
130–	66	135	8 910	144
140–	56	145	8 120	200
150–	34	155	5 270	234
160 et plus	6	165[a]	990	240
Total	240	–	32 420	–

[a] Valeurs«centrales» choisies de façon arbitraire.

Moyenne

La *moyenne approchée* est égale à la moyenne pondérée des valeurs centrales des diverses classes: elle est donc égale à 32 420/240 = 135,1 mmHg.

Médiane

La tension artérielle médiane se situe dans l'intervalle 130-140 mmHg puisque c'est là que se situent la 120^{e} et la 121^{e} observations (120 = 240/2). Les valeurs correspondantes de la tension se calculent par interpolation:

$$130+(120-78)\times 10/66,$$
$$\text{et } 130+(121-78)\times 10/66.$$

La médiane, qui est égale à la moyenne de ces deux valeurs, vaut donc 136,4 mmHg.

Plan n° 5 — leçon

Caractéristiques de dispersion

INTRODUCTION

Dans la pratique médicale, on doit classer les sujets en plusieurs catégories selon, par exemple, qu'ils sont bien portants ou malades, atteints d'une maladie donnée ou non, ont besoin ou non d'un traitement, etc. Les décisions de ce type sont prises en utilisant comme point de repère des valeurs dites «normales». La «normalité» est un concept statistique qui dépend dans une large mesure de la distribution de l'attribut considéré au sein de la population. Il est essentiel de connaître la dispersion des attributs médicaux parmi les sujets étudiés pour comprendre, utiliser et interpréter cette notion de valeur «normale». La variabilité, plus ou moins grande, des données médicales peut être récapitulée au moyen de divers paramètres de dispersion.

Objectif de la leçon

Dans la présente leçon, on se propose de définir et d'étudier divers paramètres de dispersion et de voir quelles en sont les applications, l'interprétation et les limites.

Connaissances à acquérir

A la fin de la leçon, l'étudiant devra savoir:

a) Expliquer la signification d'un paramètre de dispersion et sa place en statistique descriptive.

b) Expliquer l'emploi des termes: étendue, intervalle interquartile, variance, écart-type et coefficient de variation, en tant que paramètres de dispersion des données sanitaires.

c)† Calculer les éléments suivants, pour des données groupées ou non groupées, à l'aide de données de référence:

— étendue;

- intervalle interquartile;
- variance;
- écart-type;
- coefficient de variation.

d) Exposer les avantages et les inconvénients respectifs des cinq indices ci-dessus.

e) Choisir un paramètre approprié de dispersion pour une application donnée.

f) Etudier le concept de normalité des données sanitaires en termes de moyenne, écart-type et centiles.

Connaissances préalables requises

Tout le contenu des leçons précédentes, particulièrement au sujet de la réduction et de la présentation des données, y compris les modes de distribution. Autres points particulièrement importants: la signification et l'interprétation des caractéristiques de tendance centrale (ou de position).

Nouveaux termes et concepts

On trouvera ci-dessous la liste des nouveaux termes et concepts à étudier dans la présente leçon. S'il y a lieu, on pourra préparer un polycopié regroupant les définitions de certains d'entre eux (voir *Définitions* et *Calcul*, pp. 51-53):

coefficient de variation; distribution normale (distribution de Gauss); écart par rapport à la moyenne et somme des écarts; écart-type; étendue; intervalle interquartile et intervalle semi-interquartile (ou déviation quartile); valeurs normales; variance.

TENEUR DE LA LEÇON

Devront être traitées les questions suivantes.

Nécessité de caractéristiques de dispersion

- Emploi en pratique médicale des «valeurs normales» et des «intervalles normaux».

Exemples. Tension artérielle systolique ou diastolique, pouls, fréquence cardiaque, taille, poids, taux de cholestérol sérique, taux d'hémoglobine, etc.

- Expression de la dispersion par un paramètre unique afin de rendre plus facile la comparaison de la dispersion de différents groupes.
- Application de la dispersion comme indicateur d'homogénéité ou d'hétérogénéité des données.

Définitions

Coefficient de variation. Rapport de l'écart-type à la moyenne, exprimé sous forme d'un pourcentage (il s'agit d'un paramètre indépendant de l'échelle et de l'unité de mesure).

Distribution normale (ou distribution de Gauss). Distribution de fréquence continue, définie sur un intervalle infini et représentée par l'équation:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty.$$

La courbe de variation se caractérise par sa forme en cloche (ou en chapeau de gendarme); la moyenne, la médiane et le mode sont identiques et la distribution est parfaitement définie par la moyenne et la variance.

Ecart absolu moyen. Quotient par le nombre d'observations de la somme des écarts par rapport à un point origine convenable, ces écarts étant mesurés en valeur absolue (c'est-à-dire sans tenir compte d'un éventuel signe moins). L'écart absolu moyen est qualifié d'écart moyen par rapport à la moyenne, par rapport à la médiane ou par rapport au mode selon la nature du point choisi pour mesurer les écarts.

Ecart par rapport à la moyenne. Différence (positive ou négative) entre une observation individuelle et la moyenne du groupe. (La somme de tous ces écarts par rapport à la moyenne est égale à zéro.) On peut également calculer les écarts par rapport à la médiane ou par rapport au mode.

Ecart-type (σ). Ecart quadratique moyen dans l'hypothèse où les écarts sont calculés par rapport à la moyenne. L'écart-type est égal à la racine carrée de la variance, exprimée avec la même unité que les observations initiales.

Etendue (ou intervalle de variation). Différence entre la valeur la plus élevée et la valeur la plus faible d'une série d'observations.

Intervalle interquartile. Différence entre le premier et le troisième quartiles (Q3 - Q1). En divisant cette valeur par 2, on obtient le semi-interquartile (également appelé déviation quartile).

Valeurs normales. Valeurs considérées comme tombant dans l'intervalle de variation habituel pour une population ou un sous-groupe de population donné. L'intervalle couvert par ces valeurs est qualifié d'*intervalle normal* (ou étendue normale), défini par les limites normales.

Variance. Quotient par le nombre d'observations n (ou par $n-1$ dans le cas de la variance d'un échantillon) de la somme des carrés des écarts à la moyenne, exprimée dans une unité égale au carré de l'unité de mesure des observations initiales.

Calcul

Etendue (e). Elle se calcule en faisant la différence

$$e = x_{max} - x_{min},$$

x_{max} et x_{min} désignant respectivement la valeur la plus élevée et la valeur la plus faible d'une série d'observations.

Intervalle interquartile. Il se calcule en faisant la différence

$$Q_3 - Q_1$$

Q_1 et Q_3 désignant respectivement le premier et le troisième quartile.

On obtient de même le semi-interquartile $(Q_3 - Q_1)/2$.

Ecart absolu moyen par rapport à la moyenne. Il se calcule en faisant la somme

$$\Sigma|x_i - \bar{x}|/n,$$

formule dans laquelle x_i désigne une observation individuelle, Σ le signe «somme» qui indique qu'il faut faire la sommation pour toutes les valeurs entières de i de 1 à , $\bar{x}$ la moyenne et n la taille de l'échantillon (nombre d'observations).

Lorsqu'on remplace dans la formule ci-dessus la moyenne par la médiane ou le mode, le même calcul fournit l'écart absolu moyen par rapport à la médiane ou par rapport au mode.

Variance (S^2 pour la variance d'une population ou s^2 pour la variance d'un échantillon). Dans le cas d'une population importante, la variance se calcule au moyen de la formule:

$$S^2 = \Sigma(x_i - \bar{x})^2/n = (\Sigma x_i^2 - n\bar{x}^2)/n.$$

Si l'on a besoin d'une estimation de la variance d'une population dont on a observé un échantillon, on remplace au dénominateur l'effectif n de l'échantillon par $n-1$ de façon à obtenir une estimation sans biais. On a donc:

$$s^2 = \Sigma(x^i - \bar{x})^2/(n-1) = (\Sigma x_i^2 - n\bar{x}^2)/(n-1).$$

L'*écart-type, désigné en abrégé par la lettre grecque sigma* (σ), mais souvent aussi par les lettres S ou s, est égal à la racine carrée de la variance de l'échantillon ou de la variance de la population d'où a été extrait cet échantillon.

Coefficient de variation (C.V.). Il se calcule (lorsqu'il est exprimé en pourcentage) comme suit:

$$\text{C.V.} = (S/\bar{x}) \times 100 \text{ ou } (s/\bar{x}) \times 100.$$

Observations sur les paramètres statistiques définis ci-dessus

a) *Etendue*

- Simple à calculer.
- Facile à comprendre.
- Les valeurs extrêmes dépendent de la taille de l'échantillon.
- Ce paramètre ne tient pas compte de l'ensemble des observations, de sorte qu'il ne tient pas compte de la dispersion des observations entre les deux valeurs extrêmes.
- Il n'est guère adapté à un traitement mathématique plus raffiné.
- Il doit être utilisé parallèlement à d'autres caractéristiques de dispersion, à moins qu'on ne précise la distribution de fréquence complète et la moyenne, etc.

b) *Intervalle interquartile*

- Paramètre de dispersion simple et adapté à de multiples applications.
- Recouvre la moitié des observations, plus précisément celles qui sont situées autour du centre.
- Il ne repose pas sur l'ensemble des observations mais seulement sur la différence entre deux centiles particuliers.
- Ce paramètre permet de tenir compte des aléas de l'échantillonnage sans qu'on ait à postuler une forme particulière pour la distribution de fréquence.

— Il est important pour le choix de valeurs limites (points de troncation de la distribution globale) pour l'établissement de normes cliniques.

c) *Ecart absolu moyen par rapport à la moyenne*

— Ce paramètre est calculé à partir de la totalité des observations.

— Il est facile à comprendre.

— Il est simple à calculer.

— Il tient compte de tous les écarts en valeur absolue (c'est-à-dire sans tenir compte du sens de l'écart).

— Il ne se prête guère à un traitement mathématique plus raffiné.

d) *Variance et écart-type*

— Ces deux paramètres tiennent de l'ensemble des observations.

— Les écarts sont calculés par rapport à la moyenne.

— Les carrés des écarts sont évidemment positifs, quel que soit le sens de l'écart.

— Ce sont les deux paramètres les plus utilisés du fait des propriétés de la distribution normale théorique et de l'importance de la variance en statistiques analytiques (qui seront traitées à la leçon 6).

e) *Coefficient de variation*

— Paramètre utilisé pour comparer la dispersion relative de deux distributions.

— C'est une mesure de la dispersion des données lorsqu'on prend comme unité la valeur de la moyenne.

— Ce paramètre est indépendant de l'unité de mesure de sorte qu'il est utile pour comparer la dispersion de deux distributions dont les variables sont exprimées au moyen d'unités différentes (par exemple une taille exprimée en centimètres dans une distribution et en mètres dans une autre).

— Il tient compte de chaque valeur de la distribution.

Fixation de «valeurs normales»

La définition de «valeurs normales» permet de choisir les mesures qui conviennent dans la pratique médicale.

La dispersion est une caractéristique inhérente à toutes les mesures biomédicales qui sont à la base des décisions relatives aux programmes de santé communautaire et aux soins à dispenser à un patient déterminé. En d'autres

termes, la décision est guidée par des normes fixées au préalable. Ces normes sont souvent qualifiées de «valeurs normales» et s'appuient généralement sur des mesures effectuées au sein de groupes de population formés de sujets considérés comme «bien portants».

La décision en matière médicale exige en général deux types de valeurs normales, ponctuelles ou par intervalle. Les premières sont estimées au moyen d'une caractéristique de tendance centrale (se reporter au polycopié 4.1 où sont définies les caractéristiques de tendance centrale et de position). Un intervalle normal a pour objectif d'indiquer l'ensemble des valeurs que présente généralement un paramètre donné, dans le cas d'un groupe de population constitué de personnes en bonne santé. Certains sujets de la population présentent une valeur exceptionnellement élevée ou faible pour un paramètre donné tout en étant apparemment en parfaite santé. Ces valeurs sont qualifiées de «valeurs extrêmes». On ne peut pas considérer qu'elles soient caractéristiques du groupe de population. Il arrive donc qu'on exclue quelques observations présentant une valeur tout à fait extrême avant de calculer les valeurs normales.

La plupart des valeurs normales adoptées dans le domaine biomédical visent à faire en sorte que la valeur du paramètre correspondant tombe entre les limites ainsi fixées pour 95% d'un groupe de sujets bien portants choisis au hasard. Lorsqu'une variable répond à une distribution symétrique et unimodale, il est facile de calculer l'étendue normale en fonction de la moyenne et de l'écart-type, par exemple en utilisant les propriétés de la distribution de Gausse théorique. Par exemple, l'intervalle centré sur la moyenne et d'étendue égale à deux écarts-types ($\mu \pm \sigma$) regroupe environ 68% des sujets d'une population bien portante tandis que l'intervalle $\mu \pm 1{,}96\sigma$ en regroupe environ 95%. (On arrondit parfois, au prix d'une certaine perte de précision, à $\mu \pm 2\sigma$.) Dans le cas d'une distribution multimodale ou asymétrique, le calcul de l'étendue normale est parfois passablement complexe, bien que les mêmes principes restent valables.

Très souvent, les valeurs normales sont différentes selon la région géographique, selon le sexe ou selon le groupe d'âges. Par exemple, une tension artérielle «normale» n'est pas la même chez les deux sexes et varie également avec l'âge, selon des modalités qui ne sont pas identiques dans toutes les populations humaines. Chaque fois qu'on indique une valeur normale, il faut donc préciser à quelle population elle se rapporte.

Ecart-type et distribution normale

Lorsqu'on dit que 68% de la population est comprise dans l'intervalle $\bar{x} \pm \sigma$ et 95% dans l'intervalle $\bar{x} \pm 1{,}96\sigma$, cela n'est vrai que si l'ensemble des valeurs

est distribué normalement, autrement dit si la distribution se rapproche de la distribution théorique de probabilité dite «normale», représentée par la courbe en cloche. L'écart-type tire en grande partie son intérêt et son importance en tant que paramètre de dispersion de la place qu'il occupe dans cette distribution normale théorique.

Quand la distribution normale est rapportée à une échelle réduite (notée z), de sorte que la moyenne soit égale à 0 et l'écart-type à 1, z représente l'écart (par rapport à la moyenne) exprimé en multiple ou sous-multiple de l'écart-type de la distribution:

$$z = (x - \bar{x})/\sigma,$$

et z est désigné sous le nom d'«écart normal réduit».

PLAN DE LA LEÇON

L'exposé peut être articulé comme suit.

a) Décrire les paramètres de dispersion et exposer leur place en statistiques descriptives. Faire la différence entre un indice synthétique de tendance centrale et un indice synthétique de dispersion et expliquer en quoi leurs rôles se complètent dans l'étude de toute caractéristique présentée par un groupe de sujets (par exemple comme indicateurs d'homogénéité/hétérogénéité) et dans la comparaison de différents groupes de sujets. Expliquer que la variabilité ou dispersion peut être en rapport ou non avec la grandeur de la variable de sorte qu'il faut distinguer les indices de dispersion absolue et les indices de dispersion relative.

b) Donner les définitions et le mode de calcul des différents indices récapitulatifs de dispersion absolue qu'on rencontre couramment dans la littérature. On traitera des cas suivants:

- indices définis à partir de certains points particuliers de la distribution: étendue, intervalle interquartile, semi-interquartile;
- indices donnant une idée globale de la différence des diverses valeurs par rapport à l'un des indices de tendances centrales: écart moyen (par rapport à la moyenne ou à la médiane);
- indices donnant une idée globale du carré des différences des valeurs individuelles par rapport à la moyenne: somme des carrés; variance (ou carré de l'écart-type); écart-type (ou écart quadratique moyen, σ).

c) Appeler l'attention sur la notion d'intervalle normal, souvent fixé de façon arbitraire comme englobant 95% des valeurs centrales de la distribution de fréquence (en d'autres termes, intervalles allant du centile 2,5

au centile 97,5) et expliquer comment l'on utilise fréquemment l'écart-type pour estimer cet intervalle normal, qu'on écrit sous la forme $\bar{x} \pm 1{,}96\sigma$.

On s'attachera plus particulièrement à cette application de l'écart-type qui s'explique par les propriétés de la distribution normale théorique. On exposera la notion d'écart normal réduit z (écart par rapport à la moyenne mesuré en prenant l'écart-type comme unité) et l'on montrera, en utilisant une table où sont indiquées les valeurs de l'«aire sous la courbe», que les centiles de la distribution normale sont fonction des valeurs de z. On pourra alors se servir de la table pour résoudre des problèmes sur la proportion des valeurs de la distribution normale qui sont à l'intérieur ou à l'extérieur d'un intervalle centré sur la moyenne et d'étendue égale à un multiple ou à un sous-multiple de z (par exemple $\bar{x} \pm \sigma$, $\bar{x} \pm 1{,}96\,\sigma$). On examinera dans quel cas et pourquoi on peut (ou ne peut pas) utiliser l'écart-type de cette façon lorsqu'il s'agit de données empiriques (distribution de fréquences observées).

e) Récapituler les applications et les limites des différents paramètres de dispersion.

EXERCICES EN CLASSE

Demander aux participants de faire les exercices ci-dessous.

Exercice 1

a) Calculer l'étendue, l'intervalle interquartile, l'écart moyen (par rapport à la moyenne) en valeur absolue, la variance, l'écart-type et le coefficient de variation pour les données suivantes qui correspondent à la durée de maladie (en jours) pour 23 cas de pneumonie:

6, 7, 8, 8, 10, 11, 11, 11, 8, 10, 10, 10,
12, 12, 14, 14, 15, 15, 17, 18, 6, 5, 4.

b) En se reportant aux valeurs observées, commenter les avantages et les inconvénients de la variance, de l'écart-type et du coefficient de variation en tant que paramètre de dispersion.

Exercice 2

Le Tableau 5.1 indique, sous forme d'une distribution de fréquence, les valeurs du revenu annuel de 300 ménages, en dollars des Etats-Unis d'Amérique.

Tableau 5.1. Distribution du revenu annuel des ménages ($ E.-U.)

Revenu annuel des ménages		Fréquence
Moins de	100	2
	100–	4
	200–	9
	300–	10
	400–	22
	500–	68
	600–	85
	700–	58
	800–	25
	900–	8
	1000–	6
	1000–	2
	1200 et plus	1
Total		300

a) Calculer l'intervalle interquartile, l'écart moyen (par rapport à la moyenne) en valeur absolue, la variance, l'écart type et le coefficient de variation.

b) Commenter la distribution du revenu des ménages.

Plan n° 6 — leçon

Introduction au calcul des probabilités et aux statistiques analytiques

INTRODUCTION

La médecine n'est pas à proprement parler une science exacte: on peut la qualifier de «probabiliste», par opposition aux sciences «déterministes». Les résultats et les réponses peuvent rarement, sinon jamais, être prévus avec une parfaite certitude mais seulement avec un degré, plus ou moins important, de probabilité ou de vraisemblance.

Aussi bien la théorie des probabilités est-elle à la base des méthodes utilisées pour établir des conclusions d'ordre statistique en médecine. L'étudiant n'a pas besoin de connaître cette théorie en détail mais il faut qu'il soit familiarisé avec certains de ses concepts, principes, règles et applications de base.

Objectif de la leçon

La présente leçon vise à familiariser l'étudiant avec des probabilités, dans la mesure nécessaire pour apporter les éléments indispensables à la compréhension des développements ultérieurs sur son application dans les statistiques inductives et dans la prise de décision dans le domaine médical.

Connaissances à acquérir

A la fin de la leçon, l'étudiant devra savoir:

a) Expliquer la différence entre statistiques descriptives et statistiques analytiques et exposer les applications de ces dernières dans le domaine sanitaire.

b) Expliquer et montrer par des exemples en quoi consistent la loi des pro-

babilités totales et la loi des probabilités composées et expliquer leurs applications élémentaires en médecine.

c) Se servir des tables de probabilité (au stade actuel, tables de la loi normale et de la loi binomiale) pour résoudre des problèmes sanitaires simples.

Connaissances préalables requises

Présentation des relevés et distribution d'effectifs ou de fréquences, histogrammes et polygones des fréquences et courbes de distribution.

Indices synthétiques habituels employés en statistiques descriptives: proportions, pourcentages, centiles, moyenne et écart-type.

Nouveaux termes et concepts

On trouvera ci-dessous la liste des nouveaux termes et concepts à apprendre dans la présente leçon. Il est recommandé de distribuer avant les cours un polycopié contenant l'explication de ces nouveaux termes et concepts de façon que les étudiants puissent s'y reporter pendant les cours et ensuite (voir le polycopié 6.1 en annexe). Ces nouveaux termes et concepts sont les suivants:

attribut dichotomique; base d'échantillonnage; coefficients du binôme; distribution binomiale; distribution d'échantillonnage; distribution de probabilité; échantillon; échantillonnage; événements exclusifs; événements indépendants; fraction sondée; fréquence; induction; loi des probabilités composées; loi des probabilités totales; méthode d'échantillonnage; population; probabilité; sondage simple; taille de l'échantillon; triangle de Pascal; unité de sondage.

TENEUR DE LA LEÇON

Les points à traiter sont indiqués dans les polycopiés 6.1 et 6.2. Le professeur pourra s'en servir pour faire son cours en s'inspirant des recommandations générales ci-dessous.

PLAN DE LA LEÇON

L'exposé peut être articulé comme suit:

a) Faire un rappel rapide sur les applications des méthodes statistiques descriptives déjà enseignées et introduire la notion de statistiques inductives

en en expliquant la signification et en l'illustrant par des données médicales (par exemple, on montrera comment les critères d'«anormalité» utilisés pour poser le diagnostic reposent sur des données descriptives mais sont appliquées à de nouveaux patients.

b) Expliquer les termes et concepts définis dans la seconde partie du polycopié 6.1 («Terminologie de l'échantillonnage») en s'attachant plus particulièrement aux éléments ou aux procédés matériels d'échantillonnage et aux populations humaines (par exemple, échantillonnage de la population d'une ville dans le cadre d'une enquête de morbidité ou échantillonnage parmi les malades atteints d'une certaine affection dans le cadre d'un essai clinique).

c) Expliquer le rapport entre la probabilité et les proportions observées au sujet d'un attribut dichotomique (par exemple, la probabilité de trouver un sujet du groupe sanguin A ou la probabilité pour qu'un enfant en gestation soit de sexe masculin).

Expliquer les notions d'événements indépendants ou exclusifs et montrer comment fonctionnent la loi des probabilités totales et la loi des probabilités composées. (Par exemple, lorsque deux maladies sont «indépendantes», quelle est la probabilité de trouver un sujet atteint de ces deux maladies? Quelle est la probabilité de trouver un donneur sanguin appartenant, soit au groupe A, soit au groupe O? Se reporter au polycopié 6.1 pour la définition du qualificatif «indépendant».)

d) Introduire la notion de distribution binominale en tant que distribution de sondage.

S'agissant de données médicales dichotomiques, examiner les issues possibles, et les probabilités correspondantes, au sein d'un petit échantillon de n observations. (Par exemple, si le traitement d'une certaine maladie est efficace dans 70% des cas et si vous avez traité 5 cas (autrement dit, $p = 0{,}7$ et $n = 5$), quelles sont les chances pour que le traitement soit inefficace dans tous les cas, efficace dans un cas, etc.?) Se reporter à l'exemple avec solution du polycopié 6.2 en se servant de la table des probabilités calculées pour chacune des issues possibles (Tableau 6.1, polycopié 6.2), calculer les probabilités cumulées (dans le sens croissant ou le sens décroissant) et indiquer la probabilité pour qu'on obtienne un nombre de «réussites» inférieur ou supérieur à un nombre fixé à l'avance. (Par exemple, en reprenant les données ci-dessus, quelles sont les chances pour que le traitement soit efficace dans plus de 3 cas chez les 5 patients?)

Expliquer comment on calcule les «probabilités binomiales» en se reportant aux éléments présentés dans le polycopié 6.1 (loi binomiale et triangle de Pascal).

Donner d'autres exemples choisis dans le domaine médical et faisant intervenir des attributs dichotomiques auxquels s'applique la distribution binomiale. (Si c'est nécessaire, rédiger d'autres exemples avec solutions pour d'autres valeurs de n et p).

e) Exposer la notion générale de distributions de probabilité et de distributions des probabilités cumulées. En utilisant une table indiquant l'«aire sous la courbe de Gauss», expliquer comment on calcule la probabilité pour qu'on observe des valeurs inférieures à une valeur spécifiée du paramètre z, en utilisant des exemples médicaux. (Même si ce point a déjà été traité, par exemple à la suite des notions portant sur la moyenne et sur l'écart-type, il convient d'y revenir ici pour illustrer ce type d'application et l'imprimer dans l'esprit des participants).

f) Il est recommandé de distribuer aux participants trois polycopiés dont ils se serviront tout au long de la leçon:

i) Polycopié 6.1, couvrant les points suivants:
   - statistiques probabilistes et analytiques;
   - échantillonnage;
   - les deux lois fondamentales des probabilités;
   - distribution binomiale;
   - triangle de Pascal.

ii) Polycopié 6.2, donnant des exemples avec solutions de distribution répondant à la loi binomiale:
   - pour $n = 5$ et $p = 0{,}7$ comme éléments de référence au cours de la leçon;
   - pour $n = 10$ et $p = 0{,}5$ (ou d'autres valeurs de ces deux paramètres) à titre d'éléments de référence pendant les travaux dirigés.

iii) Un troisième polycopié donnant la table des aires sous la courbe de Gauss. Il n'est pas joint ici mais devra être établi par le professeur en se servant des éléments qu'on peut trouver dans n'importe quel manuel classique de statistiques.

EXERCICES EN CLASSE

Demander aux participants de faire les exercices ci-dessous.

Exercice 1

Pour exposer et faire comprendre la notion de distribution d'échantillonnage et montrer comment elle est régie par les lois de probabilité, faire établir par

les étudiants une distribution d'échantillonnage empirique (distribution observée) et leur faire comparer cette distribution à la distribution théorique (loi binomiale). On peut par exemple se servir pour cela de billes colorées pour représenter les personnes présentant tel ou tel attribut au sein d'une population. Donner des exemples d'attributs médicaux dichotomiques susceptibles d'être représentés par des billes de deux couleurs, par exemple noir et blanc, en prenant des exemples en génétique (anémie falciforme, groupes sanguins, etc.).

a) En utilisant une boîte contenant un grand nombre de billes de deux couleurs (par exemple noir et blanc), faire tirer par chaque étudiant un échantillon aléatoire d'effectif n donné (par exemple 5); faire un tableau contenant le nombre de billes noires (ou blanches) observées dans chaque échantillon. On obtient ainsi une distribution d'échantillonnage empirique.

b) Indiquer la proportion effective de billes noires (ou blanches) contenues dans la boîte, calculer la distribution binomiale pour un échantillon de taille n. En déduire la distribution d'échantillonnage théorique correspondant au nombre d'échantillons observés.

c) Comparer et commenter ces deux distributions — distribution observée et distribution théorique. (La validité de l'ajustement sera vérifiée plus tard lorsque les étudiants auront appris le test du chi-carré.)

Exercice 2

L'exemple ci-dessous est destiné à montrer l'application des lois des probabilités et de la distribution binomiale.

a) Fournir les valeurs de la prévalence de divers attributs (par exemple des maladies) au sein d'une population et demander quelle est la probabilité pour qu'une sujet présente diverses combinaisons de ces attributs (et, à la limite, n'en présente aucun ou les présente tous).

b) Demander quelles sont les probabilités pour qu'on observe, dans une famille de 2 ou 3 enfants, tel ou tel nombre de garçons ou de filles.

c) Présenter un exercice qui montre le rôle des énoncés probabilistes appliqués à un examen de laboratoire, par exemple sur le plan de la spécificité, de la sensibilité, de la valeur prédictive.

POLYCOPIÉ 6.1 **Nouveaux termes et concepts**

Consepts utilisés en statistiques probabilistes et en statistiques analytiques

Attribut dichotomique	Caractéristique dont la classification fait intervenir deux catégories seulement, en général, la présence ou l'absence d'un état défini (par exemple malade ou non malade; amélioration ou sans amélioration). Certaines caractéristiques sont dichotomiques par nature (par exemple homme/femme, vivant/mort), mais, même lorsque ce n'est pas le cas, les caractéristiques peuvent être «dichotomisées»; en définissant et en repérant un sous-groupe et en regroupant toutes les autres observations au sein d'un deuxième sous-groupe (qualifié de résiduel).
Déduction	Conclusion d'ordre particulier établi à partir d'une observation générale (voir *Induction*).
Dichotomie	Répartition en deux sous-classes exclusives.
Evénements exclusifs	On dit que deux événements sont exclusifs lorsqu'il est impossible qu'ils se produisent ou soient présents simultanément.
Evénements indépendants	On dit que deux événements sont indépendants lorsque la présence (survenue) ou l'absence (non survenue) ne modifie pas la probabilité ou que l'autre soit présent ou survienne.
Induction	Conclusion d'ordre général établie à partir d'une observation particulière (voir *Déduction*).
Probabilité	La probabilité peut être mesurée sur une échelle continue, allant de 0 à 1 (bornes incluses). Un événement impossible est assorti d'une probabilité égale à 0 tandis qu'un événement certain a une probablitié égale à 1. Un événement dont la probabilité est supérieure à 0,5 a davantage de chances de se produire que de ne pas se produire, etc. La notation P(A) représente la probabilité de survenue de l'événement A.

Statistiques analytiques — Méthodes statistiques qui traitent de la façon d'établir des conclusions par induction. Par opposition aux *statistiques descriptives*, les statistiques analytiques montrent comment on peut (ou ne peut pas) se servir de données concernant des groupes étudiés finis (ou des échantillons) pour en tirer, par induction, des conclusions sur la population dans son ensemble ou d'autres sous-ensembles de cette population.

Statistiques descriptives — Méthodes statistiques qui traitent de la description d'une ou plusieurs caractéristiques concernant un groupe étudié de dimension finie (voir *Statistiques analytiques*).

Terminologie de l'échantillonnage

Base d'échantillonnage — Ensemble des unités d'échantillonnage disponibles au sein desquelles on peut prélever un échantillon, par exemple liste de noms ou d'adresses ou de tous autres éléments susceptibles de constituer des unités d'échantillonnage.

Echantillon — Sous-ensemble d'une population dont certaines propriétés sont représentatives (ou doivent être considérées comme telles) de la population totale ou d'un ensemble plus vaste.

Echantillonnage — Prélèvement d'un échantillon au sein d'une population.

Fraction sondée (ou taux d'échantillonnage) — Proportion des unités d'échantillonnage à prélever dans une base d'échantillonnage donnée en vue de leur inclusion dans l'échantillon.

Méthode d'échantillonnage — Série de règles ou de spécifications à appliquer pour la constitution d'un échantillon.

Population — Tout groupe déterminé (généralement d'effectif élevé de personnes, choses ou mesures, par exemple la population étudiée, la population sondée, la population cible.

Sondage (ou échantillonnage aléatoire) simple — Méthode d'échantillonnage dans laquelle toutes les unités d'échantillonnage constituant une base donnée ont la même chance d'être prélе-

vées en vue de leur inclusion dans l'échantillon et selon laquelle tous les échantillons possibles de même taille ont *a priori* la même chance d'être constitués.

Taille de l'échantillon — Nombre de valeurs observées dans l'échantillon, généralement noté *n*.

Type d'échantillon — Manière dont les différents éléments de l'échantillon ont été prélevés ou choisis (par exemple échantillon aléatoire, échantillon par choix raisonné).

Unité d'échantillonnage — Unité retenue lors d'un échantillonnage, par exemple une persone, un ménage, une subdivision administrative. Cette unité n'est pas forcément identique à l'unité d'observation ou unité étudiée.

Univers (d'un échantillon) — Ensemble de valeurs par rapport auquel les valeurs observées sur l'échantillon constituent un échantillon aléatoire et auquel on peut valablement étendre les propriétés observées dans l'échantillon. L'univers d'un échantillon peut correspondre à une population réelle ou théorique, être fini ou infini selon le type d'échantillon et la nature des données étudiées.

Deux lois classiques de calcul des probabilités

Loi des probalités totales — Lorsque la réalisation d'un événement est tenue pour acquise quand se produit un événement quelconque faisant partie d'un groupe d'issues *exclusives l'une de l'autre*, la probabilité de cet événement est égal à la somme des probabilités de diverses issues, par exemple, dans le cas de deux issues, on a:

$$P(\text{A ou B}) = P(\text{A}) + P(\text{B}).$$

Loi des probabilités composées — Dans une série d'essais *indépendants*, la probabilité pour que se produisent tous les événements d'une série donnée est égale au produit des probabilités attachées à chacun de ces événements, autrement dit:

$$P(A \text{ et } B) = P(A) \times P(\text{B}).$$

Distribution binomiale

Elle correspond aux divers termes du développement de l'expression du binôme:

$$(p+q)^n,$$

dans laquelle n = taille de l'échantillon, p = probabilité de «réussite» et q = probabilité d'«échec» (avec $p+q = 1$).

Exemples

1. Quand $n = 2$, les termes successifs du développement de $(p+q)^2$ sont p^2, $2pq$ et q^2.
2. Quand $n = 4$, les termes du développement de $(p+q)^4$ sont p^4, $4p^3q$, $6p^2q^2$, $4pq^3$ et q^4.

Calcul des coefficients du binôme au moyen du triangle de Pascal

(*n*)										1										
1									1		1									
2								1		2		1								
3							1		3		3		1							
4						1		4		6		4		1						
5					1		5		10		10		5		1					
6				1		6		15		20		15		6		1				
7			1		7		21		35		35		21		7		1			
8		1		8		28		56		70		56		28		8		1		
9	1		9		36		84		126		126		84		36		9		1	
10	1	10		45		120		210		252		210		120		45		10		1

POLYCOPIÉ 6.2 **Exemple de distribution binomiale (exercice résolu)**

Au plus fort de la sécheresse dans une région donnée, on a estimé à 70% le nombre d'enfants de moins de 10 ans souffrant de malnutrition sévère. Si l'on choisit au hasard 5 enfants de cette tranche d'âges dans la région considérée, quelles sont les probabilités d'observer parmi eux 5, 4, 3, 2, 1 ou 0 cas de malnutrition sévère?

Tableau 6.1. Probabilités correspondant à la loi binomiale

Nombre d'enfants malnutris	Termes du développement du binôme	Probabilité
5 (tous)	p^5	0,16807
4	$5p^4q$	0,36015
3	$10p^3q^2$	0,30870
2	$10p^2q^3$	0,13230
1	$5pq^4$	0,02835
0 (aucun)	q^5	0,00243
		1,00000

Dans le tableau ci-dessus, on a : $n = 5$, $p = 0{,}7$ et $q = 0{,}3$.

Plan n° 7 — leçon

Estimations relatives à une population

INTRODUCTION

Chaque fois qu'on transpose des données observées sur un groupe fini de personnes à d'autres sujets ou à l'ensemble de la population correspondante, on se sert de données d'échantillonnage.

Les renseignements tirés de l'observation d'échantillons d'effectif restreint constituent la majeure partie, sinon la totalité, des connaissances médicales acquises sur les populations humaines. Les agents de santé font un usage constant de ce type de données, lorsqu'ils ne contribuent pas eux-mêmes à les développer. Il faut donc qu'ils soient parfaitement conscients des limites dont sont assorties ces données et les conclusions établies par induction — qu'il s'agisse de fiabilité, de précision ou de validité.

Si l'on part du principe qu'il n'existe pas deux sujets semblables, comment se peut-il que le médecin sache quelle conduite il doit tenir devant son prochain patient, qu'il n'a peut-être encore jamais vu et qui diffère à un ou plusieurs égards, des autres malades qu'il a vus jusqu'alors?

Objectif de la leçon

La présente leçon vise à fournir aux participants des connaissances théoriques et pratiques sur les erreurs d'échantillonnage et la façon dont il convient d'en tenir compte lorsqu'on tire des conclusions par induction à partir d'observations faites sur un échantillon.

Connaissances à acquérir

A la fin de la leçon, les participants devront savoir:

a) Faire la distinction entre les erreurs d'échantillonnage et les autres types d'erreur.

b) Faire la différence entre estimations ponctuelles et estimations par intervalle des indicateurs de santé (ou indicateurs sanométriques).

c)† Calculer l'erreur-type dont est entachée la moyenne de l'échantillon ou une certaine proportion, à partir des données et au moyen des formules voulues.

d) Expliquer la signification et l'utilisation des limites de confiance associées à l'estimation d'un indicateur de santé.

e) Expliquer la relation qui existe entre erreur d'échantillonnage et taille de l'échantillon, d'une part, et dispersion de la caractéristique étudiée, d'autre part.

f) Faire la différence entre un échantillonnage aléatoire et un échantillonnage raisonné.

Connaissances préalables requises

Le contenu de toutes les leçons précédentes de la série (spécialement la leçon 7 sur les concepts et termes de base en statistiques inductives et sur les distributions d'échantillonnage et leur lien avec les questions de probabilité).

Nouveaux termes et concepts

On trouvera dans le polycopié 7.1 des explications sur certains des nouveaux termes et concepts suivants:

biais (erreur systémique) d'échantillonnage; différence entre erreur d'échantillonnage et autres erreurs; échantillon naturel ou auto-sélectionné; erreur d'échantillonnage; erreur-type; estimation de la moyenne pour une population; estimation ponctuelle et estimation par intervalle; estimation d'une proportion pour une population; estimation statistique; intervalle de confiance; limites de confiance; méthode d'échantillonnage; niveau de confiance; paramètre relatif à une population; précision des estimations; qualité d'un échantillon; représentativité d'un échantillon; «statistique» relative à un échantillon; validité des estimations.

TENEUR DE LA LEÇON

Devront être traitées les questions ci-dessous.

Estimation statistique

- Grandeurs statistiques descriptives relatives à un échantillon (voir leçons 4 et 5). Les estimations relatives à un échantillon sont sans intérêt lorsqu'on dispose de données pour l'ensemble de la population.
- Notion d'échantillonnage. Raisons et circonstances d'un échantillonnage.
- Grandeurs statistiques descriptives relatives à un échantillon (voir leçons 4 et 5). Les indices établis à partir d'un échantillon représentent une estimation des paramètres relatifs à l'ensemble de la population.
- Notion d'erreur d'échantillonnage: différence, inévitable, entre la valeur d'une grandeur statistique établie à partir d'un échantillon et le paramètre correspondant pour l'ensemble de la population. L'erreur d'échantillonnage diminue avec la taille de l'échantillon. (Le cas limite correspond à celui où la taille de l'échantillon est égale à l'effectif de la population, auquel cas l'erreur d'échantillonnage est nulle).
- Validité des estimations: elle dépend du caractère représentatif de l'échantillon et non de sa taille. Elle correspond à la concordance, plus ou moins poussée, entre un estimateur et le paramètre estimé (exactitude).
- Précision d'une estimation: elle dépend de la taille de l'échantillon. La précision (fiabilité) est d'autant plus grande que l'erreur d'échantillonnage est faible.

Limites et niveaux de confiance

- Distributions d'échantillonnage: distribution des probabilités d'observer une erreur d'échantillonnage de grandeur donnée. A chaque grandeur statistique relative à un échantillon correspond une distribution d'échantillonnage. (Se reporter à la leçon 6 en ce qui concerne la distribution binomiale.)
- Estimation par intervalle: estimation d'un paramètre relatif à une population sous forme de la définition d'un intervalle ayant une probabilité fixée à l'avance de contenir la vraie valeur. L'intervalle est l'intervalle de confiance et ses limites sont les limites de confiance.
- La grandeur de l'erreur d'échantillonnage détermine la précision de l'estimation par intervalle. Par exemple, connaissant la moyenne ($\bar{x}$) et l'erreur-type (e_q) pour un échantillon, les limites de confiance à 95% de moyenne estimative pour l'ensemble de la population sont $\bar{x} \pm 1{,}96\, e_q$.

— Calcul de l'erreur-type pour une moyenne: si s représente l'estimateur de l'écart-type de la population, établi en fonction des valeurs observées sur l'échantillon (voir leçon 5), l'erreur-type correspondant à la moyenne d'un échantillon d'effectif n est égale à:

$$s/\sqrt{n}.$$

Pour une population de taille n distribuée selon une loi binomiale de probabilité p (voir leçon 6) et si l'on observe a individus présentant la caractéristique étudiée, l'erreur-type associée à l'estimation de p, c'est-à-dire a/n, vaut:

$$\sqrt{pq/n}.$$

Taille de l'échantillon

Dans le cas d'un échantillonnage aléatoire simple et si l'on désigne par d le niveau de confiance, on peut évaluer la précision z au moyen de la formule:

$$z = \frac{d}{e_q}.$$

Dans le cas d'un intervalle de confiance à 95%, z doit être égal à 1,96 (voir p. 54). Comme l'erreur-type dépend de n, on peut calculer la valeur de n nécessaire pour atteindre le niveau de confiance choisi.

Dans le cas de l'estimation de la moyenne pour une population où l'on a vu que l'erreur-type valait $s/\sqrt{n}$, on a de façon générale:

$$z = \frac{d}{s/\sqrt{n}}.$$

Dans ces conditions, la taille minimale nécessaire pour l'échantillon vaut:

$$n = z^2 s^2/d^2.$$

Dans le cas d'une distribution binomiale où l'erreur-type est égale à $\sqrt{pq/n}$ (voir plus haut), ce qui donne:

$$n = z^2 pq/d^2.$$

Ces résultats ne sont valables que si l'échantillon est extrait d'une population très nombreuse (théoriquement infinie) et que le taux d'échantillonnage est très faible.

Si l'échantillon est extrait d'une population finie de taille N, la taille minimale de l'échantillon est égale à =

$$n = z^2s^2/(d^2+z^2s^2/N),$$

pour l'estimation de la moyenne et à:

$$n = z^2pq/(d^2+z^2pq/N),$$

pour l'estimation du paramètre p de la distribution binomiale.

Si n_o représente l'effectif de l'échantillon extrait d'une population infinie, l'effectif de l'échantillon extrait d'une population finie est égal à:

$$n = n_o/(1+n_o/N).$$

(Voir un exemple avec solution dans la leçon 15, pp. 168-169.)

STRUCTURE DE LA LEÇON

L'exposé peut être articulé comme suit:

a) Cadre général d'une estimation statistique: nécessité d'estimer les paramètres relatifs à une population à partir de grandeurs statistiques concernant un échantillon; répercussion de l'erreur d'échantillonnage sur la fiabilité des estimations; notions d'estimation ponctuelle et d'estimation par intervalle; considérations de validité et de précision d'une estimation statistique.

b) Notions de limites de confiance et de niveaux de confiance; rapports entre distributions d'échantillonnage, unités de confiance et niveau de confiance.

c) Relation entre limites de confiance et erreur-type: signification de l'erreur-type; bases du calcul et définition des limites de confiance en fonction de l'erreur-type.

d) Calcul de l'erreur-type associée à une moyenne ou à une proportion: exemples tirés de la littérature en vue d'illustrer le mode de calcul de ces paramètres et leur utilisation dans les estimations statistiques.

e) Montrer comment on peut utiliser les équations indiquées pour estimer la taille nécessaire pour un échantillon compte tenu de la précision souhaitée, exprimée sous forme d'un intervalle de confiance.

f) Présentation d'exemples illustrant l'estimation de valeurs anthropométriques normales pour une population, du poids de naissance moyen des enfants nés en milieu hospitalier et de la prévalence d'une maladie observée lors d'une enquête de morbidité.

EXERCICES EN CLASSE

Demander aux étudiants de faire les exercices ci-dessous.

Exercice 1

Présenter les données correspondant à un paramètre ou à un attribut relatif, par exemple, à un groupe de malades hospitalisés pour l'affection X ou un groupe d'étudiants en médecine.

Les étudiants devront essayer de définir la population à laquelle on peut généraliser les données ainsi observées (autrement dit l'univers de l'échantillon) et d'examiner les biais possibles lors de la généralisation, par exemple, à toutes les personnes atteintes de la maladie X ou à toutes les personnes du pays considéré d'âge semblable à celui des étudiants en médecine.

Les étudiants devront ensuite calculer l'erreur-type et les limites de confiance à 95% et faire des commentaires sur leur intérêt dans le cas de données relatives à un échantillon naturel (auto-sélectionné).

Exercice 2

Compte tenu de l'objectif fixé pour un sondage à l'étude, calculer la taille de l'échantillon à observer. On fournira aux étudiants les données nécessaires pour le calcul.

Exercice 3

En présence d'un schéma thérapeutique réputé efficace contre une certaine maladie, demander aux étudiants d'indiquer entre quelles limites doit se situer le nombre de guérisons quand on applique ce schéma à un nombre déterminé de patients pour qu'on puisse effectivement parler d'efficacité.

Exercice 4

Le compte rendu d'un essai clinique fait état d'un taux de guérison de $p\%$ lorsqu'on administre le traitement T à n patients atteints de la maladie M. Faire calculer par les étudiants l'erreur-type du taux de guérison et les limites de confiance à 95%. Ils devront par ailleurs noter la façon dont les cas ont été choisis en vue de l'essai clinique et voir dans quelle mesure le taux de guérison déterminé dans l'échantillon est valablement représentatif du taux de guérison qu'on observerait si l'on administrait le même traitement à tous les patients atteints de la maladie M. Les étudiants devront faire la différence entre les éléments qui déterminent la validité et ceux qui déterminent la précision du taux de guérison estimatif.

POLYCOPIÉ 7.1 **Définitions des nouveaux termes et concepts**

Distribution d'échantillonnage — Distribution des probabilités d'observer une erreur d'échantillonnage d'une grandeur déterminée pour des raisons strictement aléatoires dans le cas d'un échantillon de taille donnée et pour une «statistique» particulière. On peut établir cette distribution expérimentalement en portant les valeurs obtenues pour la même statistique lorsqu'on prélève au hasard dans la même population plusieurs échantillons de même taille. On peut également établir la distribution de façon théorique (par exemple distribution d'échantillonnage répondant à la loi normale ou à la loi binomiale). Chaque valeur de la statistique représente un élément d'une distribution d'échantillonnage, à savoir la distribution des valeurs que peut *a priori* prendre cette statistique dans différents échantillons de même taille tirés au hasard du même univers.

Erreur d'échantillonnage — Différence uniquement due au hasard entre la valeur d'une statistique et la valeur du paramètre correspondant au niveau de la population (par exemple différence entre la valeur de la moyenne pour un échantillon aléatoire et de la moyenne pour l'univers correspondant). L'erreur d'échantillonnage est impossible à éviter ou à éliminer totalement de sorte qu'il faut toujours en tenir compte lorsqu'on tire des conclusions, par induction et déduction, de statistiques observées sur un échantillon. On peut réduire cette erreur en augmentant la taille de l'échantillon.

Erreur-type (e_q) — Ecart-type d'une distribution d'échantillonnage. L'*erreur-type* d'une statistique correspond à l'*écart-type* de la distribution d'échantillonnage correspondante.

Estimation par intervalle

Estimation d'un paramètre relatif à une population sous forme d'un intervalle de valeurs probables.

Estimation ponctuelle

Estimation d'un paramètre relatif à une population par une valeur unique, la plus probable. Une estimation ponctuelle est généralement fournie par une statistique. En soi, ce type d'estimation néglige l'erreur d'échantillonnage.

Limites de confiance

Limites (bornes) supérieure et inférieure de l'intervalle des valeurs les plus probables lorsqu'on procède à une estimation par intervalle. L'intervalle lui-même est désigné sous le nom d'intervalle de confiance. L'expression «limites de confiance» s'explique par le fait que ces limites sont fixées par référence à un niveau de confiance déterminé ou habituel qui indique la probabilité pour que le paramètre estimé tombe à l'intérieur de ces limites. Par exemple, les limites de confiance à 95% déterminent un intervalle qui a 95 chances sur 100 de contenir le paramètre estimé. Souvent, on peut calculer les limites de confiance à partir de l'*erreur-type.*

Niveau de confiance

Il est généralement choisi égal à 95% (soit 0,95), mais on peut choisir une valeur plus faible ou plus élevée si on le souhaite.

Paramètre relatif à une population

Indice descriptif dont la valeur se rapporte à l'ensemble d'une population et non à un échantillon extrait de cette population (par exemple une moyenne ou une proportion).

Précision d'une estimation

Inverse de l'erreur-type de l'estimation. Plus l'erreur d'échantillonnage est *a priori* faible, autrement dit plus l'intervalle de confiance est d'étendue limitée, plus la précision est élevée. On peut donc exprimer la précision en fonction de l'intervalle de confiance ou de l'erreur-type.

Statistique (ou fonction discriminante)

Indice descriptif dont la valeur se calcule à partir des observations faites sur un échantillon (par exemple moyenne ou proportion relative à un échantillon).

Validité d'une estimation

Degré de concordance entre une estimation et le paramètre estimé correspondant. Elle ne dépend pas de la taille de l'échantillon mais de la représentativité de celui-ci. Autrement dit, elle dépend du type de la nature de l'échantillon, de son mode de sélection et de l'exactitude tant des données qui sont à la base des calculs que du calcul lui-même.

Plan n° 8 — leçon

Signification statistique d'une différence

INTRODUCTION

Le hasard peut expliquer des différences entre les groupes étudiés de sorte que chaque fois que l'on observe une différence, on doit se poser la question de sa signification statistique, c'est-à-dire de la faible probabilité pour que cette différence soit le seul fait du hasard. Cela explique qu'il soit fait mention de tests de signification dans la littérature médicale chaque fois qu'on recourt aux méthodes statistiques. On les rencontre dans les articles scientifiques où l'on rend compte d'essais cliniques, d'études épidémiologiques ou d'autres recherches dans le domaine de la santé.

Objectif de la leçon

On cherchera dans ce qui suit à faire comprendre aux étudiants en quoi consiste les tests de signification et quelles sont leurs applications et leur rôle en statistiques inductives. L'accent est mis sur les applications et l'interprétation davantage que sur la théorie et la méthodologie des tests.

Connaissances à acquérir

A la fin de la leçon, l'étudiant devra savoir:

a) Situer et expliquer la notion de signification statistique.

b) Expliquer à quel moment et pour quelle raison il est indispensable d'effectuer un test de signification.

c) Faire la différence entre risque de première espèce et risque de seconde espèce dans les tests d'hypothèse.

d) Expliquer pourquoi l'expression «non improbable» n'est pas forcément synonyme de «probable» dans le cadre des statistiques inductives.

e) Exposer les résultats possibles d'un test de signification et leur interprétation respective compte tenu du contexte.

f) Distinguer les situations dans lesquelles on peut appliquer le test du chi-carré et le test du *t* de Student.

g) Distinguer les cas dans lesquels on utilise un test unilatéral ou un test bilatéral.

h) Expliquer la nature des conclusions possibles sur les «causes» d'une différence significative ou non.

i) Distinguer signification sur le plan statistique et signification sur le plan biologique.

j)† Appliquer un test du chi-carré en cas de besoin, avec l'aide des données de référence.

k)† Appliquer le test du *t* de Student en cas de besoin, avec l'aide des données de référence.

Connaissances préalables requises

L'étudiant devra avoir suivi tous les cours précédents de la série. Il est souhaitable qu'au début de la présente leçon le professeur insiste sur la nécessité, pour les étudiants, d'avoir acquis les connaissances correspondant aux leçons 6 et 7 et, notamment, d'avoir assimilé les notions d'erreur d'échantillonnage et de distribution d'échantillonnage. Il fera bien de s'en assurer.

Nouveau termes et concepts

Sont abordés dans la présente leçon les nouveaux termes et concepts suivants (qui sont définis dans le polycopié 8.1):

acceptation d'une hypothèse; degrés de liberté; hypothèse nulle; «nature aléatoire improbable»; niveau de signification; «non improbable» et «probable»; notation $P<0{,}05$, $p<0{,}01$, etc.; probabilité pour qu'une différence constatée soit le seul fait du hasard; rejet d'une hypothèse; risques de première et de seconde espèce; signification statistique; statistique d'un test (*z*, *t*, chi-carré); test d'hypothèse; test unilatéral ou bilatéral.

TENEUR DE LA LEÇON

Le professeur peut établir son plan en utilisant les définitions et explications données au polycopié 8.1 pour les nouveaux termes et concepts. Il doit également se reporter aux indications données sur la structure de la leçon.

PLAN DE LA LEÇON

Tout au long du cours, on fera abondamment usage d'exemples pris dans la littérature pour illustrer la place de la signification statistique dans l'interprétation des données et des conclusions qu'on en tire.

On donnera aux étudiants, dans la mesure jugée utile et nécessaire, des exemples avec solutions montrant comment on pratique effectivement un test de signification, conformément à la nature des connaissances qu'on souhaite leur faire acquérir. On trouvera dans les polycopiés 8.2 et 8.3, en annexes, des exemples d'applications possibles du test du χ^2 (chi-carré) et du test t. La leçon peut être articulée comme suit:

a) Présentation de la notion de signification statistique et de ses conditions d'application, dans les cas suivants:

- différences, par exemple, entre le nombre prévu et le nombre observé de survenues d'un certain événement ou différence entre les moyennes, ou les proportions, correspondant à différents échantillons;
- formulation et test de l'hypothèse nulle;
- probabilité pour qu'une différence supérieure ou égale à une valeur donnée soit le simple fait du hasard. Illustrer ce phénomène par rapport à la distribution d'échantillonnage théorique;
- sens de la différence et conséquences pour les test unilatéraux ou bilatéraux;
- risque d'erreur lorsqu'on rejette ou accepte une hypothèse; erreurs de première et de seconde espèce.

b) Introduction de la notion de niveau de signification.

- Jusqu'à combien doit tomber la probabilité P d'un événement pour qu'il puisse être considéré comme «improbable» et, par conséquent, qu'on puisse rejeter l'hypothèse nulle et décider que la différence observée est statistiquement significative?
- Indiquer les niveaux classiques de signification, à savoir $P<0{,}05$: significatif; $P<0{,}01$: très significatif; $P\geqslant 0{,}05$: non significatif.

c) Expliquer le rôle des tests de signification et les conséquences qui découlent de leur résultat. Examiner les causes possibles d'une différence observée, à savoir le hasard (hypothèse nulle), le facteur étudié, d'autres facteurs «réel», des facteurs «de confusion», par exemple un biais ou la non-comparabilité.

Un test de signification tient uniquement compte de l'aspect aléatoire; expliquer comment l'on tient compte des autres causes possibles expliquant la différence observée. Insister sur la différence entre signification *statistique* et signification *médicale.*

Expliquer comment on interprète le résultat d'un test de signification, en envisageant les diverses solutions ci-dessous:

«rejet» ou «impossibilité de rejeter» ou «acceptation» de l'hypothèse nulle; «improbable» ou «non improbable» ou «probable» que la différence soit due au hasard; différence «significative» ou «notable» ou «importante».

d) Exposé de la méthodologie des divers tests de signification.

Il existe de nombreux types de test dont le domaine d'application dépend de la nature des données et des différences étudiées. Ceux dont on se sert le plus souvent sont le test z, le test t et le test du χ^2 (chi-carré). Il faut faire appliquer au moins l'un de ces tests par les étudiants de façon qu'ils puissent assimiler les notions et principes en cause et sachent:

- choisir le test approprié à utiliser;
- calculer la statistique du test;
- évaluer sa valeur compte tenu de la distribution d'échantillonnage théorique correspondante, en indiquant quelle est la probabilité pour que cette valeur soit le simple fait du hasard (dans le cas où l'hypothèse nulle est exacte);
- décider si la différence est significative ou non et, dans l'affirmative, indiquer le niveau de signification.

Les exemples résolus des polycopiés 8.2 et 8.3 constituent des modèles dont les étudiants pourront s'inspirer lorsqu'ils auront besoin d'appliquer ces tests.

EXERCICES EN CLASSE

Des exercices à faire en classe sont prévus pour que les élèves puissent s'entraîner à l'application des tests de signification, à l'interprétation des résultats obtenus et qu'ils aient l'occasion d'en étudier le rôle dans la réalisation des objectifs de l'étude.

Les étudiants doivent formuler des hypothèses à tester et effectuer toute une série de tests de signification sur des ensembles convenables de données. Les données peuvent être artificielles mais il serait préférable qu'elles correspondent à une situation réelle et soient recueillies, soit par observation directe par les étudiants, soit par le dépouillement des publications actuelles dans le domaine de la médecine ou des domaines connexes. Les exemples pris dans la littérature ont l'avantage de constituer des études de cas et de bien montrer dans quelle mesure le test de signification a contribué aux conclusions de l'auteur.

Pour faire ressortir l'enchaînement du présent cours et faire la synthèse des différentes leçons, on peut également pratiquer des tests de signification sur des données utilisées précédemment dans d'autres exercices quand c'est possible. Par exemple, on peut se servir des distributions d'échantillonnage de la leçon 6, qui étaient conformes à la loi binomiale, pour voir dans quelle mesure il y a bonne concordance entre la distribution théorique et la distribution observée. De même, on peut reprendre les données qui ont servi au calcul de moyennes et d'écarts-types pour voir si la différence des valeurs observées d'un sous-groupe à l'autre est ou non significative.

POLYCOPIÉ 8.1 **Définitions des nouveaux termes et concepts**

Hypothèse nulle

Hypothèse de «l'absence de toute différence» ou, de façon plus correcte, l'hypothèse pour que la différence observée soit exclusivement une erreur d'échantillonnage, c'est-à-dire qu'elle soit le fait du hasard. Dans un test de signification, on formule une certaine hypothèse, qualifiée d'hypothèse nulle, qui sert de base au calcul de la probabilité pour qu'une différence soit de nature purement aléatoire. Quand la différence n'est pas significative, on accepte l'hypothèse nulle; quand elle est significative, on rejette l'hypothèse nulle en faveur d'autres hypothèses (dites «alternatives») sur les causes de la différence. A noter qu'on ne *démontre* jamais que l'hypothèse nulle est rigoureusement exacte ou fausse, mais simplement qu'on la *rejette* ou qu'on l'*accepte* au niveau de signification retenu, à savoir 0,05, 0,01, etc.

Niveau de signification

Probabilité pour qu'une différence constatée soit le simple fait du hasard, une valeur plus faible étant considérée comme suffisamment «improbable» pour que l'on considère que la différence est statistiquement significative (valeur classiquement retenue: 0,05).

Test unilatéral ou bilatéral

Lorsqu'on ne précise pas le sens de la différence qui fait l'objet d'un test de signification (par exemple lorsqu'on ne distingue pas les cas $X_1<X_2$ ou $X_1>X_2$), on tient compte, dans le test de signification, des probabilités correspondant aux deux extrémités (queues) de la distribution d'échantillonnage: en d'autres termes, il faut se servir d'un test bilatéral. Quand le sens de la différence qui fait l'objet d'un test de signification est précisé au départ (par exemple lorsqu'on compare le cas $X_1<X_2$, mais non le cas inverse $X_1>X_2$, à l'hypothèse nulle $X_1=X_2$), il convient d'utiliser un test unilatéral puisqu'on s'intéresse uniquement à la probabilité $P(X_1<X_2)$ et non à la probabilité $P(X_1>X_2)$.

POLYCOPIÉ 8.2 **Exemples d'application du test du chi-carré (avec solutions)**

Exemple 1

L'exemple ci-dessous correspond à l'application du test du chi-carré avec 2 degrés de liberté[a] à des données présentées selon un tableau de contingences 2×3 (Tableau 8.1).

Fréquences observées (O)

Dans une enquête effectuée dans un village, on a enregistré l'approvisionnement en eau de 124 ménages. L'examen des dossiers de morbidité du centre de santé du village pour la période de trois mois précédant l'enquête, a permis de repérer les membres des ménages ayant souffert de diarrhée à cette époque (Tableau 8.1).

Tableau 8.1. Antécédents diarrhéiques sur une période de 3 mois

	Nombre de ménages pour chaque mode d'approvisionnement en eau			
Etat	rivière	puits	eau courante	Total
Aucun épisode diarrhéique	39	14	12	65
Episodes diarrhéiques	49	6	4	59
Total	88	20	16	124

Fréquences attendues (A)

Les fréquences attendues (A) pour chaque case du tableau (dans le cas de l'hypothèse nulle, c'est-à-dire de l'absence d'association entre la source d'approvisionnement en eau et les épisodes diarrhétiques) sont les suivants:

46,13	10,48	8,39	65,00
41,87	9,52	7,61	59,00
88,00	20,00	16,00	124,00

[a] Dans le cas d'un tableau de contingence, le nombre de degrés de liberté est égal au nombre de cases qu'on peut remplir de façon arbitraire, pour des valeurs données des totaux marginaux. Pour un tableau à k lignes et à h colonnes, le nombre de degrés de liberté est égal à $(k-1) \times (h-1)$.

Les valeurs de la statistique $(O-A)^2/A$ sont les suivantes pour les diverses cases:

1,102	1,179	1,556
1,214	1,299	1,715

$\chi^2 = \Sigma(O-A)^2\ A$ avec $(k-1)\ (h-1)$ degrés de liberté
$= 8{,}06$ avec 2 degrés de liberté $(0{,}01 < P < 0{,}05)$.

Ainsi, l'hypothèse nulle serait rejetée au niveau de 5%. Ces données indiquent par conséquent qu'il existe un *lien* entre les épisodes diarrhéiques et la source d'eau utilisée dans le village (avec des proportions de 56%, 30% et 25% d'épisodes diarrhéiques dans les ménages qui prélèvent leur eau à la rivière, dans un puits ou un robinet, respectivement). Des investigations plus approfondies seraient nécessaires pour démontrer l'existence d'une relation causale.

Exemple 2

L'exemple qui suit illustre l'utilisation d'une formule spéciale pour l'application du chi-carré aux données d'un tableau de contingence 2×2 (tableau doublement dichotomique).

Les données du Tableau 8.2 proviennent d'une étude sur le rapport entre l'utilisation de «la pilule» par la mère et l'existence d'un ictère chez le nourrisson au sein.

Tableau 8.2. Tableau de contingence 2×2 (doublement dichotomique)

Prise de la pilule	Nourrissons		Total
	Avec ictère	Sans ictère	
Oui	33 (a)	24 (b)	57 ($a+b$)
Non	14 (c)	45 (d)	59 ($c+d$)
Total	47 ($a+c$)	69 ($b+d$)	116 (n)

$$\chi^2 \text{ «corrigé»} = \frac{n(|ad-bc|-0.5n)^2}{(a+b)(c+d)(a+c)(b+d)} = 12{,}66\ (P<0{,}001).$$

$$\chi^2 \text{ «non corrigé»} = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = 14{,}04\ (P<0{,}001).$$

POLYCOPIÉ 8.3 **Exemple d'application du test du** t de Student (avec solution)

Les données ci-dessous proviennent d'une étude qui portait sur la comparaison de la plombémie moyenne (en mg de plomb par 100 g de sang) chez un groupe d'ouvriers employés dans une fabrique de piles (ouvriers professionnellement exposés) et chez un groupe d'ouvriers employés dans une usine textile (ouvriers non exposés professionnellement).

Ouvriers de la fabrique de piles	Ouvriers de l'usine textile
0,082	0,040
0,080	0,035
0,079	0,036
0,069	0,039
0,085	0,040
0,090	0,046
0,086	0,040

Ouvriers de la fabrique de piles	Ouvriers de l'usine textile
$\Sigma X_1 = 0{,}571$	$\Sigma X_2 = 0{,}276$
$\Sigma X_1^2 = 0{,}046847$	$\Sigma X_2^2 = 0{,}010957$
$\Sigma x_1^2 = 0{,}0002697143$	$\Sigma x_2^2 = 0{,}0000757143$
$s_1^2 = 0{,}0000449524$	$s_2^2 = 0{,}0000126190$
$s_1 = 0{,}0067047$	$s_2 = 0{,}0035523$
$\bar{X}_1 = 0{,}08157$	$\bar{X}_2 = 0{,}03943$
$n_1 = 7$	$n_2 = 7$

avec $x_1 = X_1 - \bar{X}_1$ et $x_2 = X_2 - \bar{X}_2$. On en déduit:

$$s^2 \text{ (ensemble des observations)} = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

$$= 0{,}000028786;$$

et

$$e_q(d) = s\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$= 0{,}002868,$$

les indices 1 et 2 se rapportant respectivement aux ouvriers de la fabrique de piles et à ceux de l'usine textile.

L'hypothèse nulle (H_0) correspond à l'identité de la plombémie moyenne chez les ouvriers des deux industries. Cette hypothèse nécessite l'application d'un test bilatéral. On a

$$d = \bar{X}_1 - \bar{X}_2$$
$$= 0{,}042414.$$

Pour contrôler l'hypothèse nulle, on compare ces différences au chiffre zéro, avec

$t = d/e_q(d)$ *et* $(n_1 + n_2 - 2)$ degrés de liberté, soit:

$= 14{,}7$ et 12 degrés de liberté; $P < 0{,}001$.

Par conséquent, il faut rejeter l'hypothèse nulle.

Plan n° 9 — leçon

Association, corrélation et régression

INTRODUCTION

L'idée de relation causale détermine en grande partie la décision médicale, tant dans le domaine préventif que dans le domaine thérapeutique. Mais, le plus souvent, les preuves d'une relation de cause à effet sont de nature statistique, s'agissant des sciences médicales de sorte que les étudiants en médecine doivent comprendre les fondements statistiques de la connaissance de ces relations: c'est à ce prix seulement qu'ils pourront apprécier les limites des conclusions dont il est fait état dans la littérature et évaluer aussi leur propre expérience de façon plus rationnelle, quantitative et objective.

Objectif de la leçon

La présente leçon vise à apporter aux étudiants une certaine connaissance des observations statistiques témoignant de relations entre différentes caractéristiques ou différents événements au sein d'une population de façon qu'ils puissent appliquer et interpréter les méthodes et indices statistiques dont on se sert pour décrire et exprimer quantitativement ces relations.

Connaissances à acquérir

A la fin de la leçon, l'étudiant devra savoir:

a) Donner des exemples des types de questions qui se posent en matière de santé ou de médecine et dont la réponse passe par l'analyse d'une association ou d'une corrélation statistique.

b) Expliquer la signification de l'indépendance statistique entre caractéristiques ou événements ainsi que l'intensité et le sens d'une relation statistique.

c) Faire la différence entre une relation statistique et une relation causale.

d) Expliquer la notion de coefficient d'association ou de corrélation (linéaire) et les propriétés classiques de ce coefficient.

e) Expliquer ce qui fait qu'une relation observée est ou n'est pas statistiquement significative.

f) Expliquer pourquoi il arrive qu'une relation statistiquement significative soir dépourvue d'utilité ou d'importance.

g) Utiliser correctement le coefficient Q d'association.

h) Interpréter la nature de l'association de données présentées sous forme d'un diagramme de dispersion.

i) Utiliser correctement le coefficient r de corrélation.

j) Expliquer la notion de régression linéaire et ses modalités d'application.

k)† Calculer le coefficient Q à l'aide des données de référence si nécessaire.

l)† Calculer le coefficient r à l'aide des données de référence si nécessaire.

m)† Etablir l'équation et faire le tracé d'une droite de régression linéaire à l'aide des données de référence.

Connaissances préalables requises

Le contenu de toutes les leçons précédentes de la série.

Nouveaux termes et concepts

Devront être étudiés dans la présente leçon les nouveax termes et concepts suivants:

analyse bivariate (à deux variables); association entre attributs dichotomiques (tableau de contingence 2×2 — ou doublement dichotomique — et coefficient Q d'association); coefficient d'association ou de corrélation; comparaison des notions de relation statistique et de relation causale; corrélation variables continues — tableau de corrélation, diagramme de dispersion, type de relations (linéaire/non linéaire, coefficient de corrélation linéaire r); intensité d'une relation (complète, parfaite; partielle, imparfaite); régression linéaire — variable dépendante et variable indépendante, estimation d'une variable à partir de l'autre, courbe de régression; sens d'une relation (fonction croissante, corrélation positive; fonc-

tion décroissante, corrélation négative); signification statistique de l'association ou de la corrélation.

TENEUR DE LA LEÇON

Relations

Le cours doit porter sur les deux points suivants:

a) Signification et importance de l'étude des relations dans le domaine médical, des points de vue suivants:
- contribution au savoir médical;
- rôle dans la décision médicale;
- problèmes posés par le caractère multifactoriel des causes ou des effets;
- problèmes posés par la variabilité des réponses ou des manifestations.

b) Méthodes d'étude des relations et implications du résultat:
- analyse bivariate, tableau à entrées multiples;
- traitement des différents types de données ou échelles de mesure;
- relations entre des moyennes, ou proportions, relatives à différents groupes;
- différences entre relation statistique et relation causale;
- sens et intensité d'une relation;
- établissement des relations statistiques.

Régression linéaire

On étudiera la notion de régression linéaire et ses applications en expliquant en quoi consiste les variables «dépendantes» et les variables «indépendantes». On présentera la droite de régression linéaire en fournissant des précisions sur les points suivants:
- l'équation $Y=bX+c$;
- le coefficient de régression ou pente b de la droite, fourni par l'expression $\Sigma(X-\bar{X})\ (Y-\bar{Y})/\Sigma(X-\bar{X})^2$;
- l'ordonnée à l'origine (donnée par $\bar{Y}-b\bar{X}$);
- le calcul des coefficients b et c;
- le tracé de la droite de régression sur le diagramme de dispersion.

PLAN DE LA LEÇON

Après un rappel rapide des éléments des leçons précédentes qui sont essentiels à la compréhension de la présente leçon, on pourra articuler le cours comme suit:

a) Etudier la signification et l'importance de l'étude des relations dans le domaine médical, en se servant d'exemples publiés.

b) Décrire l'étude des relations entre deux variables (ou caractéristiques), en faisant la différence entre relation causale et relation statistique. Faire ressortir les conséquences qui découlent de variables ou facteurs indépendants (variables explicatives) dans l'interprétation des corrélations et associations. Par exemple, quand on interprète les différences observées en ce qui concerne la prévalence des maladies diarrhéiques de l'enfance (variable expliquée) d'une communauté (ou d'une région) à l'autre, quelles sont les implications de facteurs comme l'approvisionnement en eau, les conditions de vie, le degré d'instruction de la mère, le revenu, l'accès à des établissements de soins fonctionnels, etc. (variables/facteurs explicatifs)?

c) Etudier l'emploi des indices d'association dans le cas d'attributs dichotomiques ou dans celui de variables continues (coefficient Q d'association et coefficient r de corrélation).

d) Illustrer ces divers points par des exemples:

- tirés des facteurs de risque liés à une maladie donnée, pour faire ressortir la nature de l'observation statistique d'une association et de la notion de risque relatif;
- tirés d'essais cliniques pour mettre en évidence la nature de la preuve statistique de relation entre traitement et réponse;
- tirés de la littérature pour montrer dans quel contexte on se sert des analyses de régression et de corrélation;
- de décisions ou actions quotidiennes dont la justification réside dans la croyance, fondée ou erronée, à une association (par exemple croyances traditionnelles et superstitions);
- tirés de la presse journalière, spécialement au sujet de questions controversées retenant actuellement l'attention;
- de nature numérique de façon à présenter des situations dans lesquelles le coefficient d'association ou de corrélation a une valeur négative ou positive, ou une valeur absolue proche de 0 ou de 1.

EXERCICES EN CLASSE

Demander aux participants de faire les exercices suivants.

Exercice 1

Procéder à une enquête parmi les étudiants présents de façon à réunir des données sur deux attributs dichotomiques, choisis par les étudiants eux-mêmes, en vue de rechercher la présence possible d'associations entre ces attributs. Cela fait, les étudiants devront:

a) Présenter les données sous forme d'un tableau de contingence 2×2.

b) Calculer le coefficient Q d'association.

c) Rechercher s'il existe ou non une association statistiquement significative.

d) Commenter les résultats.

Exercice 2

Les étudiants devront proposer deux variables (par exemple le poids et la taille) faciles à mesurer en classe, en vue d'étudier une éventuelle corrélation entre ces variables. Les mesures devront être effectuées sur un petit échantillon des personnes présentes de façon à obtenir, par exemple, 10 couples de valeur en vue de l'analyse. Les étudiants devront ensuite:

a) Reporter les données sur un diagramme de dispersion.

b) Calculer le coefficient r de corrélation.

c) Etablir l'équation de régression linéaire de Y en X (Y désignant la variable dépendante et X la variable indépendante) et tracer la droite de régression sur le diagramme de dispersion.

d) Rechercher s'il existe ou non une corrélation statistiquement significative.

e) Commenter les résultats obtenus.

Au lieu de demander aux étudiants d'établir leurs propres données pour faire les exercices ci-dessus, on pourra prendre des données convenables dans la littérature ou tirés des propres travaux de l'enseignant.

POLYCOPIÉ 9.1 **Exemples de valeurs du coefficient Q d'association**

Exemple 1

		B +	B −	
A	+	10	0	10
	−	0	10	10
		10	10	20

$Q = (100 - 0)/(100 + 0) = 1{,}0.$

Example 2

		B +	B −	
A	+	8	2	10
	−	0	10	10
		8	12	20

$Q = (80 - 0)/(80 + 0) = 1{,}0.$

Exemple 3

		B +	B −	
A	+	0	8	8
	−	8	4	12
		8	12	20

$Q = (0 - 64)/(0 + 64) = -1{,}0.$

Example 4

		B +	B −	
A	+	6	4	10
	−	2	8	10
		8	12	20

$Q = (48 - 8)/(48 + 8) = 0{,}71.$

Exemple 5

		B +	B −	
A	+	4	6	10
	−	8	2	10
		12	8	20

$Q = (8 - 48)/(8 + 48) = -0{,}71.$

Example 6

		B +	B −	
A	+	60	40	100
	−	2	8	10
		62	48	110

$Q = (480 - 80)/(480 + 80) = 0{,}71.$

Exemple 7

		B +	B −	
A	+	3	7	10
	−	3	7	10
		6	10	20

$Q = (21 - 21)/(21 + 21) = 0.$

POLYCOPIÉ 9.2 **Exemple de calcul du coefficient r et de détermination de l'équation de régression**

N° du couple	X	Y	X^2	Y^2	XY
1	42	46			
2	66	35			
3	56	62			
4	80	60			
5	37	54			
6	70	74			
7	51	33			
8	38	21			
9	68	52			
10	26	34			
	534	471	31 270	24 567	26 593

Etape 1

$n = 10,$
$\bar{X} = 534/10 = 53{,}4,$
$\bar{Y} = 471/10 = 47{,}1.$

Etape 2

Posons $x = (X - \bar{X})$ et $y = (Y - \bar{Y})$.
$\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/n = 31\,270 - (534)^2/10 = 2\,754{,}4,$
$\Sigma y^2 = \Sigma Y^2 - (\Sigma Y)^2/n = 24\,567 - (471)^2/10 = 2\,382{,}9,$
$\Sigma xy = \Sigma XY - (\Sigma X)(\Sigma Y)/n = 26\,593 - (534) \times (471)/10 = 1\,441{,}6.$

Etape 3

$$r = \Sigma xy/\sqrt{(\Sigma x^2)(\Sigma y^2)}$$
$$= 1\,441{,}6/\sqrt{((2.754{,}4) \times (2\,382{,}9))}$$
$$= 0{,}5627.$$

Etape 4

Pour la régression de Y en X (estimation de Y en fonction de X):

$$b = \Sigma xy/\Sigma x^2 = 1\,441{,}6/2\,754{,}4 = 0{,}5234,$$
$$c = \bar{Y} - b\bar{X} = 47{,}1 - (0{,}5234) \times (53{,}4) = 19{,}15.$$

L'équation de la droite de régression (de Y en X) est donc:

$$Y = 0{,}5234\,X + 19{,}15.$$

Plan n° 10 — séminaire

Critique d'un article scientifique

INTRODUCTION

Les médecins diplômés puisent largement dans la littérature médicale pour se tenir au courant des dernières acquisitions dans leur domaine. Ce faisant, ils se trouvent en présence de publications de qualité et d'intégrité scientifiques extrêmement variables de sorte qu'ils doivent s'en remettre à leur propre jugement pour apprécier rationnellement le bien-fondé et la fiabilité des données présentées, la validité des conclusions établies sur cette base et les recommandations formulées par les auteurs. L'acquisition de cette capacité constitue un objectif global important pour un cours de biostatistiques destiné à des étudiants en médecine.

Les étudiants doivent être à même de repérer les points forts ou les faiblesses de tout article et de se faire une idée équilibrée, impartiale et objective de sa qualité. Ils ne doivent pas prendre pour argent comptant toutes les données qui sont publiées ni se cantonner dans une critique négative systématique.

Objectif du séminaire

Le présent séminaire est destiné à donner l'occasion aux étudiants d'appliquer les connaissances acquises dans les leçons précédentes de ce cours de biostatistiques à l'évaluation des données statistiques présentées dans une publication scientifique et d'illustrer la façon dont il convient de soumettre un article de ce type à une analyse critique.

Connaissances à acquérir

A la fin du séminaire, l'étudiant devra savoir:

a) Enumérer les principaux éléments qu'il convient d'examiner lorsqu'on procède à l'analyse critique d'une publication scientifique.

b) Montrer comment il faut s'y prendre avec chacun de ces éléments en prenant l'exemple d'une publication existante.

CONDUITE DU SÉMINAIRE

Le séminaire doit être axé sur les étudiants.

L'enseignant choisira une publication scientifique adaptée dans la littérature actuelle, portant sur un sujet intéressant et à la portée des étudiants. Environ une semaine avant la date prévue pour le séminaire, il distribuera une copie de la publication à chaque membre du séminaire, ainsi qu'un document indiquant la façon de procéder à une analyse critique.

On demandera à deux ou trois étudiants de préparer un bref exposé théorique et de conduire les débats lors du séminaire. Tous les autres étudiants auront à procéder à l'avance à l'analyse critique du document, conformément aux instructions reçues, et devront, si l'enseignant souhaite les noter, remettre leurs conclusions par écrit avant le séminaire. La remise d'une critique par écrit peut être souhaitable si le groupe d'étudiants est trop important pour que chacun participe efficacement à la discussion.

Bien que la marche à suivre pour la critique d'un rapport médical soit indiquée dans le polycopié 10.1, il n'existe en réalité aucune façon de procéder qui soit idéale pour tous les types de rapports médicaux ou d'articles scientifiques: l'enseignant pourra donc s'il le souhaite utiliser un plan différent ou modifier le plan proposé ici de façon à l'adapter au document particulier retenu pour le séminaire.

Lors du choix de ce document, on pourra attacher davantage d'importance au bon choix du sujet et de la méthodologie qu'à la présence éventuelle de défauts importants dans l'étude. Il y a autant, sinon plus, à apprendre de l'analyse critique détaillée d'une publication de qualité que d'une publication présentant à l'évidence de graves défauts.

TENEUR DU SÉMINAIRE

Pour l'essentiel, le séminaire sera articulé conformément aux rubriques et aux éléments indiqués dans le polycopié remis aux étudiants, moyennant certaines modifications, le cas échéant, pour l'adapter à l'article qui fera l'objet d'une discussion.

La tâche qui incombe à l'enseignant est de faire en sorte que les étudiants soient objectifs, sans idée préconçue, réalistes et, de façon générale, fassent preuve de rationalité dans leur évaluation de l'article. On encouragera les critiques constructives de préférence aux critiques destructives. Si les étudiants fournissent des suggestions ou recommandations, il faudra que l'enseignant veille à ce qu'elles soient également soumises à une analyse critique au lieu d'être simplement mises en avant, de façon superficielle, en tant que solutions de rechange. Par exemple, chaque fois qu'on déclare que le cher-

cheur aurait dû s'y prendre de telle façon de préférence à telle autre, il est indispensable de comparer les avantages comme les inconvénients des deux solutions et de se demander ce qui a poussé le chercheur à procéder comme il est indiqué dans le rapport.

Il importe que l'enseignant soit suffisamment familiarisé avec l'étude pour pouvoir appeler l'attention des personnes présentes sur les points saillants qu'ils auraient laissé échapper.

L'objectif primordial de la critique est d'évaluer la validité de la ou des conclusions formulées par l'auteur ou les auteurs de l'article; l'enseignant pourra avoir à intervenir au cours du séminaire, en faisant des commentaires ou en posant des questions pour éviter que la discussion s'écarte trop de cet objectif.

POLYCOPIÉ 10.1 **Plan de l'analyse critique d'un article médical**[a]

Objectifs ou hypothèses

a) Quels sont les objectifs de l'étude ou les questions auxquels on cherche une réponse?

b) Quelle est la population à laquelle les chercheurs se proposent d'appliquer leurs conclusions?

Nature de l'étude

a) De quel type était l'étude: expérience, observations planifiées, ou analyse de dossiers?

b) Comment l'échantillon a-t-il été choisi? Existe-t-il des biais éventuels pouvant rendre l'échantillon atypique ou non représentatif? Dans l'affirmative, comment en a-t-on tenu compte?

c) Quelle est la nature du groupe témoin ou de la norme de comparaison?

Observations

a) A-t-on défini la terminologie employée, qu'il s'agisse de critères diagnostiques, des mesures effectuées ou des critères d'issue?

b) La méthode de classification ou de mesure adoptée était-elle uniforme pour tous les sujets et adaptée aux objectifs de l'investigation? Existe-t-il des biais possibles au niveau des mesures et, dans l'affirmative, quelles dispositions a-t-on prises pour en tenir compte?

c) Les observations sont-elles fiables et reproductibles?

Présentation des observations

a) Les conclusions sont-elles présentées clairement, objectivement et de façon suffisamment détaillée pour que le lecteur puisse lui-même les apprécier?

b) Les observations présentent-elles une cohérence interne: exactitude des totaux, possibilité de rapprocher les différents tableaux, etc.?

---

[a] COLTON, T *Statistics in medicine*, Boston, Little, Brown & Co., 1974.

Analyse

a) Les données sont-elles susceptibles de faire l'objet d'une analyse statistique? Dans l'affirmative, les méthodes d'analyse statistique sont-elles appropriées à la source et à la nature des données, et l'analyse est-elle correctement effectuée et interprétée?

b) L'analyse est-elle suffisamment poussée pour qu'on puisse voir si les «différences significatives» ne sont pas en réalité dues à l'absence de comparabilité des groupes en ce qui concerne leur distribution par sexe ou par âge, leurs caractéristiques cliniques ou d'autres paramètres pertinents?

Conclusions

a) Quelles sont les conclusions justifiées par les observations et celles qui ne le sont pas?

b) Les conclusions sont-elles en rapport avec les questions posées au départ par les chercheurs?

Propositions constructives

Supposez que vous êtes en train de planifier une investigation en vue de répondre aux questions posées dans la présente étude. Si ces questions n'ont pas été clairement énoncées par les auteurs, donnez-en une formulation appropriée. Proposez une organisation pratique de l'investigation, des critères à retenir pour les observations et la nature de l'analyse à pratiquer de façon à réunir des données fiables et valables en rapport avec les questions étudiées.