Biostatistique

connaissances de base

R. Giorgi

Coordonnateur

M. Fieschi, H. Chaudet, J. Gaudart, B. Giusiano, J. Gouvernet, J. Mancini

ont collaboré à l'élaboration de ce document

Table des matières

Avant-propos

Ce document est destiné aux étudiants inscrits au Master Santé Publique dans les spécialité « Expertise et Ingénierie des Systèmes d'Information en Santé », « Méthodes Quantitatives et Econométriques pour la Recherche en Santé », « Santé Publique, Société Développement » dans une Unité d'Enseignement traitant de biostatistique ou d'épidémiologie.

L'objectif de ce document est de rappeler les concepts de base en biostatistique :

- type de variable observé ;
- bases probabilistes ;
- modélisation de la variabilité et de l'incertitude ;
- principe de l'estimation statistique ;
- principe d'un test statistique.

Certaines de ces notions seront développées et approfondies lors des enseignements spécifiques.

Chapitre 1 Introduction générale aux statistiques

Généralités

Population, échantillon

La méthode statistique, en général, a pour but de dégager certaines propriétés d'un ensemble de mesures (ou d'observations) ou de décrire cet ensemble (appelé population pour des raisons historiques).

Une **population** peut être tout aussi bien un groupe d'êtres humains, un ensemble d'objets ; tous ces éléments ayant en commun un attribut ou une propriété qui caractérise cet ensemble d'éléments (exemple : les individus de sexe masculin).

Généralement, le statisticien n'étudie pas le caractère sur l'ensemble de la population mais sur un **échantillon** extrait de la population, pour plusieurs raisons, entre autres :

- La taille de la population peut être très importante et le coût de l'enquête serait trop important ;
- L'accès à tous les individus de la population est matériellement impossible ;
- L'étude du caractère peut détruire les éléments de la population.

Le nombre d'éléments constituant l'échantillon est appelé **l'effectif** ou la **taille de l'échantillon**.

Un bon échantillon doit constituer une image réduite de l'ensemble de la population dont on veut étudier un caractère bien défini. Dans le cas contraire, on dit que l'échantillon est **biaisé**.

Le choix de l'échantillon, le recueil des données nécessaires à l'étude que l'on se propose de conduire, constituent la partie fondamentale, la plus longue, de l'étude.

Afin de généraliser les résultats obtenus sur l'échantillon, on désire que celui-ci représente le mieux possible la **population cible** c'est à dire celle sur laquelle porte l'étude.

Échantillonnage

Comment choisir un échantillon pour qu'il soit représentatif ?

Il existe plusieurs techniques d'échantillonnage :

Tirage au hasard

Un échantillon ne doit en aucun cas être choisi par commodité. Afin de disposer d'un **échantillon représentatif**, il faut le constituer d'une manière « **aléatoire** » : on peut pour cela procéder à un véritable **tirage au sort** ou bien utiliser des tables de nombres aléatoires qui ont été construites à cet effet.

On peut constituer un échantillon par un tirage au hasard dans toute la population ou bien par des procédés plus complexes comme la stratification.

Stratification

On subdivise la population en sous groupes (ou **strates**) et on choisit ensuite l'échantillon en tirant au sort dans chacune des strates. Chaque strate peut être représentée en fonction de son importance dans la population.

> *Exemples :*
>
> *1. Si l'on veut faire une enquête épidémiologique sur l'hypertension artérielle, on pourra constituer un échantillon qui sera un modèle réduit de la population étudiée. En stratifiant de telle sorte qu'il respecte les mêmes proportions que la composition de la population quant aux catégories socioprofessionnelles, aux tranches d'âges, au sexe …*
>
> *2. Dans un essai thérapeutique d'un traitement anticancéreux, on pourra définir les strates en tenant compte des facteurs pronostiques tels que : taille de la tumeur, extension loco-régionale, métastase à distance, …*

Il faut remarquer qu'il n'est pas toujours facile de prélever un bon échantillon. Le prélèvement de l'échantillon doit être fait au **hasard**. Nous allons voir sur un exemple les difficultés qui peuvent être rencontrées dans le choix des échantillons :

> *Exemple :*
>
> *On se propose d'étudier le pourcentage de décès dans la population française des sujets atteints d'un infarctus du myocarde.*
>
> *On peut constituer un échantillon en observant les décès des malades qui ont été hospitalisés dans un service hospitalier donné. Le biais introduit, si la population « cible » est la population de tous les français, est évident.*
>
> *En effet, le service hospitalier a un recrutement particulier et une renommée telle qu'il hérite, peut-être, de malades plus graves, ou d'une catégorie sociale dont le genre de vie, l'alimentation, l'âge, …, sont des facteurs pronostiques qui peuvent modifier l'issue de la phase aiguë.*
>
> *Un échantillon représentatif de la population française atteinte d'un infarctus du myocarde pourrait être obtenu par tirage au sort sur tous les cas d'infarctus du myocarde recensés en France. Toutefois on ne les connaît pas tous et il est toujours possible d'introduire un biais.*

Problème de l'estimation

Il s'agit d'évaluer un paramètre (une caractéristique) sur un échantillon pour pouvoir estimer ce paramètre pour la population entière. Le problème de l'estimation est développé plus loin.

> *Exemple :*
>
> *Évaluation, à partir de la mesure de la glycémie pratiquée sur un échantillon de sujets sains ayant entre 20 et 40 ans, de la valeur moyenne de la glycémie pour tous les sujets sains de cette tranche d'âge.*
>
> *Si l'on veut que cette estimation soit aussi précise que possible, il est nécessaire que l'échantillon soit aussi représentatif que possible de la population.*

Les tests statistiques

Il s'agit de tirer des conclusions sur la population à partir de l'étude d'un ou plusieurs caractères observés sur les individus d'un ou de plusieurs échantillons issus de cette population.

Ce problème inclut celui de la comparaison de caractéristiques (une ou plusieurs) issues de 2 ou plusieurs populations (comparer la glycémie moyenne des sujets urbains et des sujets ruraux). Il trouve sa solution dans les tests statistiques qui sont des tests d'hypothèses. Ils permettent de faire des inférences statistiques.

Les hypothèses que l'on veut tester sont imposées par construction du test (développé plus loin).

Chapitre 2 Statistique descriptive

Buts de la statistique descriptive

Toute série d'observations comporte un certain nombre de données relatives à un ou plusieurs caractères ou encore **variables**. Le but des statistiques descriptives est de **décrire un ensemble d'observations** à l'aide de quelques éléments caractéristiques.

> *Exemple : la taille des français adultes.*
>
> *Dans ce cas, les mesures seront nombreuses, le tableau des données, c'est-à-dire la liste des tailles de tous les sujets, ne donnera, au premier abord, aucun renseignement clair. Grâce aux statistiques descriptives, on caractérisera cet ensemble d'individus par un moyen simple qui réduira le nombre de données. Par exemple, si on s'intéresse à la taille, on procédera au calcul de la valeur moyenne de la taille. De cette façon, il est certain que l'on perd de l'information, mais on gagne en commodité de présentation des données.*

Pour présenter les données, le premier travail consiste donc à rassembler et à présenter clairement les observations. Plusieurs cas sont à envisager suivant le type des données recueillies.

Les différents types de données

Données de type qualitatif

Un caractère est **qualitatif** s'il peut se présenter sous plusieurs aspects ou suivant plusieurs modalités. Ces données donnent lieu à des dénombrements.

> *Exemples :*
>
> *Le sexe, la couleur des yeux, l'efficacité ou la non efficacité d'un traitement, la nature des cellules d'un tissu, le groupe sanguin, …*

On est amené à définir des catégories ou **classes exclusives** correspondant aux différentes modalités du caractère observé, puis à déterminer à quelle classe appartient chaque individu. Un individu appartient à une classe et une seule.

Données de type ordinal

Il est possible qu'il existe entre les diverses classes une relation d'ordre, telle que par

exemple : plus grave que …, de meilleur pronostic que …

Le caractère observé est alors de type **ordinal**.

Exemple :

Classification en stades 1, 2, 3, 4 des patients atteints de la maladie de Hodgkin.

Les malades au stade 2 sont plus gravement atteints que ceux qui sont classés au stade 1, …

Le mécanisme de base reste le même : on affecte chaque individu à une classe et une seule. Toutefois, notons qu'il existe un ordre sur les classes.

Données de type quantitatif

Une variable **quantitative** prend pour valeur un nombre résultant de la mesure, avec une unité, du caractère chez chaque individu. La mesure est telle qu'une même différence entre des valeurs observées a toujours la même signification.

Exemple de la mesure de la taille :

Soit 4 individus A, B, C, D dont les tailles sont exprimées en centimètres :

A = 175 cm ; B = 180 cm ; C = 165 cm ; D = 170 cm.

(On peut dire que 180 - 175 = 170 - 165 = 5 cm).

Un caractère quantitatif est discret ou continu suivant qu'il est susceptible de prendre des valeurs isolées ou bien toutes les valeurs de son intervalle de variation.

***Exemples de caractères quantitatifs discontinus** (ou discrets) : nombre d'enfants dans une fratrie, nombre de cellules par mm^3, …*

***Exemples de caractères quantitatifs continus** : tension artérielle, glycémie, …*

Caractérisation des données qualitatives et ordinales unidimensionnelles

Fréquence absolue et tableau des effectifs

La **fréquence absolue** est le nombre d'individus par classe. Ce dénombrement donne lieu à une représentation des données sous forme de tableau.

Exemple :

On a dénombré sur un ensemble de 180 sujets, les individus qui appartenaient aux différents groupes sanguins (Tableau 2.1).

A+	***A-***	***B+***	***B-***	***AB+***	***AB-***	***O+***	***O-***
80	*10*	*20*	*5*	*5*	*2*	*50*	*8*

Tableau 2.1 : Description de l'échantillon des groupes sanguins.

Sur les classes ainsi formées, seules les opérations suivantes sont permises : réaliser des classes disjointes à partir d'une seule classe, ou bien regrouper certaines classes. La seule relation qui puisse être utilisée sur ces données est la relation d'appartenance à une même classe.

Exemple (suite) :

Sur l'exemple ci-dessus, on pourrait regrouper les classes correspondant aux rhésus + ou -, ou ignorer le rhésus pour former les groupes A, B, AB, O (Tableau 2.2).

A	***B***	***AB***	***O***
90	*25*	*7*	*58*

Tableau 2.2 : Description de l'échantillon des groupes sanguins sans facteur rhésus.

Fréquences relatives

On peut définir les **fréquences relatives** qui sont, pour chaque classe, le rapport de son effectif au nombre total d'individus de la série des mesures.

La somme des fréquences relatives est égale à 1.

Parfois, les résultats sont exprimés en pourcentage, chacune des fréquences relatives étant multipliée par 100 (Tableau 2.3).

A	**B**	**AB**	**O**
50	14	4	32

Tableau 2.3 : Fréquences relatives (exprimées en pourcentage et arrondies à l'unité).

Fréquences cumulées (relatives et absolues)

Les **fréquences cumulées** sont utilisées pour les **données ordinales** qui présentent des classes ordonnées.

Exemple :

Sur un échantillon de 500 malades cancéreux, on a noté le stade de la maladie. On peut résumer ou présenter ces données par des fréquences relatives. Les résultats obtenus sont présentés par la figure 2.4.

Stade	*Nombre de malades*	*Fréquence relative (%)*	*Fréquence relative cumulée (%)*
1	*350*	*70*	*70*
2	*110*	*22*	*92*
3	*30*	*6*	*98*
4	*10*	*2*	*100*

Tableau 2.4 : Répartition du stade de la maladie.

Cette présentation permet de dire, par exemple, que 92% des sujets examinés ont un stade inférieur ou égal à 2.

Diagramme « camembert »

On peut représenter les effectifs absolus ou relatifs des classes par des secteurs de cercle dont la surface est proportionnelle à l'effectif.

Le diagramme « camembert » ainsi construit est bien adapté à la représentation des données qualitatives « pures » (exemple Tableau 2.5 et Figure 2.1).

Yeux	*Marron*	*Vert*	*Bleu*	*Noir*
Effectif	*50*	*10*	*28*	*12*

Tableau 2.5 : Couleur des yeux dans un échantillon de 100 sujets.

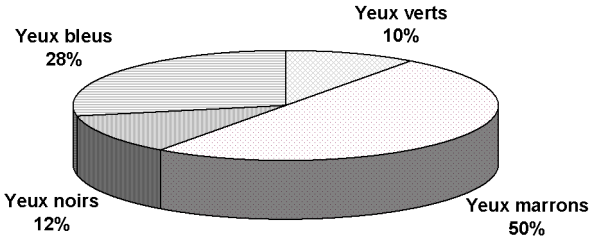

Figure 2.1 : Diagramme en camembert.

Diagramme en bâtons, mode

Diagramme en bâtons

Pour les **données ordinales,** on peut également représenter les fréquences absolues, relatives ou cumulées par un diagramme en bâtons.

Exemple :

L'exemple de l'échantillon des 500 cancéreux dont on a noté le stade est

représenté sur la Figure 2.2.

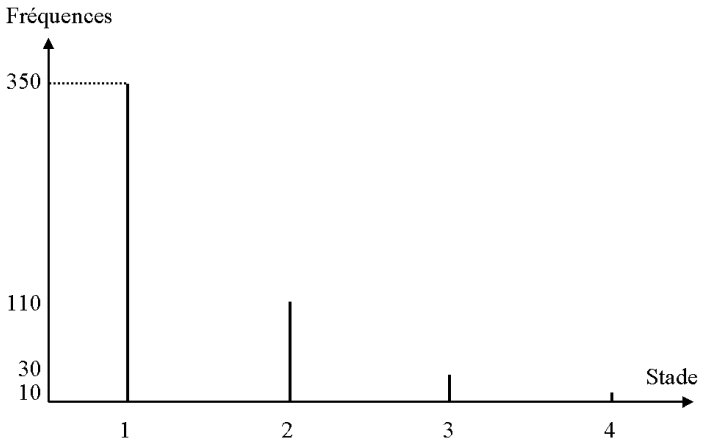

Figure 2.2 : Diagramme en bâtons des stades de la maladie.

Mode

Sur l'exemple de la Figure 2.2, la classe caractérisée par le stade 1 est la classe qui contient le plus grand nombre de sujets ; c'est le **mode** ou classe modale. Le mode est la classe (catégorie) qui offre la plus grande fréquence

Dans le cas de **variables ordinales**, si les données montrent plusieurs classes d'effectifs supérieurs aux effectifs des classes adjacentes, on dit que le diagramme représente une distribution **multimodale** : bi-modale, tri-modale, … Dans le cas contraire, on dit que la distribution est **uni-modale**.

Caractérisation des données qualitatives à deux dimensions

Les modalités de deux variables qualitatives permettent de constituer des classes exclusives, auxquelles nous affectons chaque observation, classiquement représentées sous la forme d'un tableau appelé **tableau de contingence**.

Exemple :

Dans un échantillon de 200 sujets on a relevé la présence ou l'absence d'un signe clinique S et d'une maladie M (Tableau 2.6).

	M+	***M-***	***Total***
S+	*90*	*30*	*120*
S-	*30*	*50*	*80*
Total	*120*	*80*	*200*

Tableau 2.6 : Tableau de contingence (les malades présentant la maladie sont dénombrés dans la colonne M+, les autres dans la colonne M-).

Ce tableau comporte deux parties :

- Les effectifs dénombrés pour chacune des modalités, pour chacun des deux

caractères étudiés ;

- Les effectifs de chaque modalité d'un caractère, quelles que soient les modalités de l'autre caractère. Ces effectifs sont situés dans la dernière colonne et la dernière ligne.

La dernière ligne et la dernière colonne sont appelées : les « **marginales** », (marge ligne et marge colonne) ou encore « distributions marginales ».

Caractérisation des données quantitatives à une dimension

Généralités

Nous avons déjà vu que les variables quantitatives peuvent être de deux types : variables discontinues (ou discrètes) et variables continues.

Dans le cas des variables discontinues, il est possible de représenter les données par un diagramme en bâtons, comme dans le cas de données ordinales.

Dans tous les cas, on peut diviser l'intervalle de variation de la variable en un certain nombre de classes et l'on dénombre toutes les mesures à l'intérieur de chaque classe.

Exemple :

Soit la série de mesures représentant les âges de 20 individus, rangées par ordre croissant :

3, 5, 6, 7, 8, 11, 15, 20, 21, 22, 23, 23, 23, 30, 31, 32, 35, 36, 40, 45.

On peut décider de déterminer des classes d'âge de 10 ans en 10 ans[1] *: 0 - 10 ans, 10 - 20 ans, 20 - 30 ans, 30 - 40 ans, 40 - 50 ans. On transforme ainsi la série qui peut se représenter dans le tableau des fréquences (Tableau 2.7).*

Classe	***Effectif / classe***
0 - 10 ans	*5*
10 - 20 ans	*2*
20 - 30 ans	*6*
30 - 40 ans	*5*
40 - 50 ans	*2*

Tableau 2.7 : Effectifs par classe.

Histogramme

Les **données quantitatives continues** peuvent être représentées par un

[1] Nous adoptons la convention suivante : la borne supérieure de l'intervalle est exclue.

histogramme.

Dans un système d'axes on se propose de représenter le Tableau 2.7. On porte sur l'axe des abscisses les extrémités de chaque classe et pour chacune d'elles on construit un rectangle dont la base est le segment limité aux extrémités de la classe et **la surface est proportionnelle à l'effectif de la classe**. La surface limitée par la ligne polygonale obtenue en bordant la partie supérieure de l'ensemble des rectangles s'appelle l'**histogramme** (Figure 2.3).

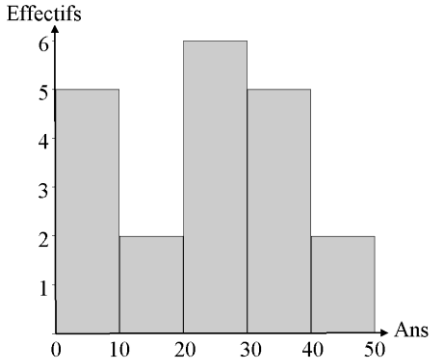

Figure 2.3 : Histogramme.

Un tel histogramme est tracé en respectant deux règles :

- L'échelle sur l'axe des abscisses est identique pour tous les intervalles de classes ;
- La surface de chacun des rectangles est proportionnelle au nombre d'individus de la classe.

La deuxième règle se simplifie si les intervalles de classe ont tous la même largeur. Cette simplification est très souvent utilisée. En effet quand les intervalles de classe sont de même largeur, la hauteur du rectangle est proportionnelle à l'effectif, ce qui facilite la lecture de l'histogramme.

Le contour polygonal joignant les milieux des bases supérieures des rectangles s'appelle le **polygone des fréquences** (Figure 2.4).

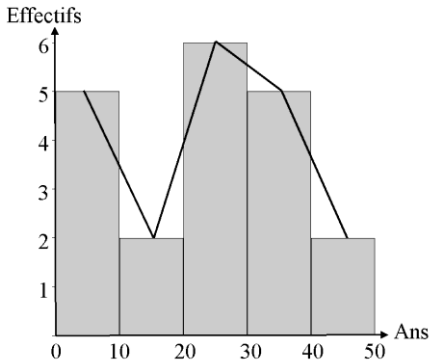

Figure 2.4 : Histogramme et polygone des fréquences.

Si on augmente le nombre des classes, de même largeur, recouvrant l'étendue de l'échantillon, l'intervalle de chaque classe devenant très petit, on peut admettre, à condition que la population soit « infinie », que l'histogramme et le polygone des

fréquences se « rapprochent », et que leur limite commune est une courbe continue (Figure 2.5).

Figure 2.5 : Courbe de fréquences et distribution de fréquences.

Cette courbe est dite « courbe des fréquences ». Si l'on rapporte la fréquence absolue de chaque classe à l'effectif total de l'échantillon, on obtient la fréquence relative par classe.

L'ensemble des classes affectées de leur fréquence constitue **une distribution de fréquences**.

Quand le nombre de classes tend vers l'infini, le polygone des fréquences devient une ligne continue : **la courbe des fréquences**.

Paramètres statistiques décrivant un ensemble de mesures quantitatives

En général, il est intéressant de présenter de façon simple et abrégée les caractéristiques principales de l'ensemble des mesures qui ont été effectuées sur un échantillon ou une population.

On utilise dans ce but quelques grandeurs numériques appelées paramètres de la distribution qui peuvent être réparties en deux catégories.

On distingue :

- **Les paramètres de position : moyenne, médiane, mode, quantiles ;**
- **Les paramètres de dispersion : variance, écart-type, intervalle inter-quartile.**

Ces paramètres font partie des grandeurs statistiques que l'on nomme parfois « statistiques ». On dira par exemple que le calcul de la moyenne est le calcul d'une statistique.

Paramètres de tendance centrale ou de position

Ces paramètres définissent l'ordre de grandeur des mesures effectuées, c'est-à-dire de l'ensemble des mesures de la distribution. Ce sont les valeurs autour desquelles se groupent les différentes mesures effectuées sur l'échantillon.

La moyenne

La valeur centrale la plus utilisée est la moyenne arithmétique des mesures, c'est-à-dire le rapport de la somme des mesures au nombre de mesures effectuées.

On peut caractériser la tendance centrale par la moyenne notée

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \left(\frac{x_i}{n} \right)$$

La moyenne s'exprime dans les mêmes unités que les valeurs observées.

Usuellement on note $\{x_1, x_2, \ldots, x_n\}$ la série de mesures, $\bar{x}$ la moyenne et n l'effectif.

Exemple :

Considérons la série de mesures constituée par les poids de 5 individus (poids exprimés en kilogrammes) : 70,0 ; 68,5 ; 72,5 ; 73,0 ; 76,0. La moyenne est égale à 72 kg.

La médiane

La médiane est la valeur qui laisse de part et d'autre un nombre égal d'observations. C'est donc un nombre de même nature et de même unité que les valeurs observées.

Pour déterminer la médiane d'une série de nombres, il est nécessaire d'ordonner cette série de mesures.

Exemple (suite) :

Dans l'exemple précédent, il faut ordonner les poids : 68,5 ; 70 ; 72,5 ; 73,0 ; 76,0.

La médiane est égale à 72,5 Kg car il y a autant de mesures inférieures à 72,5 que de mesures supérieurs à 72,5.

Deux cas peuvent se présenter :

- Si n est impair ($n = 2k + 1$), la médiane est la valeur de la mesure qui se situe au milieu de la série de mesures ordonnées : c'est x_{k+1}.
- Si n est pair ($n = 2k$), on appelle médiane toute valeur comprise entre x_k et x_{k+1}. En effet, il n'y a pas de valeur observée qui soit au milieu de la série de mesures. En général, on prend pour valeur de la médiane $(x_{k+1} + x_k)/2$.

Remarques :

1. La médiane est moins influencée que la moyenne arithmétique par les valeurs extrêmes.

   En effet, si dans la série précédente, le plus petit des poids, c'est-à-dire 68,5 kg est remplacé par 55 kg, la moyenne est influencée alors que la médiane reste identique.

2. La médiane peut aussi être utilisée dans le cas des données ordinales, puisque sa détermination se base sur l'ordre des données.

Le mode

Le mode, encore appelé **valeur dominante**, est la valeur de la variable dont la fréquence est maximale.

Si les données sont affectées à des classes, on parle de **classe modale**. La classe modale est celle dont la fréquence est maximale. C'est, dans le cas d'une courbe continue des fréquences, l'abscisse du point d'ordonnée maximale.

Si la distribution de fréquences est symétrique et unimodale (Figure 2.6), moyenne, médiane et mode sont confondus.

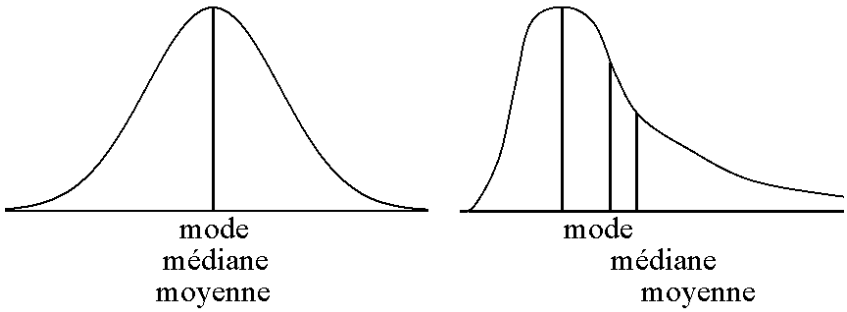

Figure 2.6 : Distributions symétrique et dissymétrique.

Dans certaines distributions, il peut n'y avoir qu'un petit nombre d'observations dans le voisinage de la moyenne ou de la médiane (Figure 2.7).

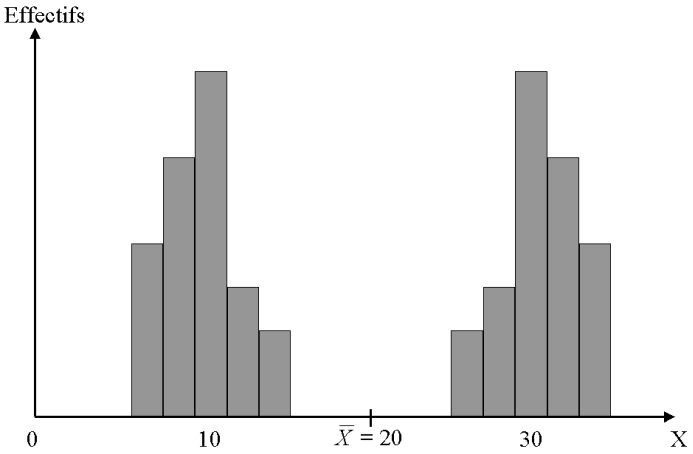

Figure 2.7 : Moyenne et dispersion.

On remarque, sur la Figure 2.7 qu'aucune valeur de X ne se trouve près de la moyenne et de la médiane.

Une distribution peut avoir plusieurs classes modales, elle est dite alors : bimodale, trimodale, …, plurimodale (plusieurs classes dont les effectifs sont grands, séparées par des classe à effectifs faibles). Une telle distribution traduit généralement un échantillon d'individus hétérogènes.

Remarque

Propriété de la moyenne :

Soit la série $\{x_1, x_2, \ldots, x_n\}$ de moyenne $\bar{x}$ et deux constantes a et b.

Considérons la série : $\{y_1, \ldots, y_n\}$ telle que :

$y_1 = ax_1 + b$, $y_2 = ax_2 + b$, …, $y_n = ax_n + b$,

soit : {ax_1 + b ; … ; ax_n + b}.

La moyenne de cette nouvelle série $\bar{y}$ est égale à : $\bar{y} = a\bar{x} + b$.

Les quantiles

Les quantiles sont les **valeurs de la variable** qui divisent l'échantillon ordonné en groupes d'effectifs égaux.

Les quantiles portent des noms différents selon le nombre de groupes souhaités : quartiles pour 4 groupes, déciles pour 10 groupes, percentiles pour cent groupes.

Les quartiles : Pour séparer les valeurs de la variable en quatre groupes d'effectifs égaux, il faut trois valeurs appelées quartiles. Les quartiles sont les valeurs Q1, Q2 et Q3 de la série X qui partagent l'effectif total, après l'avoir ordonné, en 4 classes de même effectif.

Entre la valeur minimum de la série et le premier quartile Q1, on retrouve un quart des observations. Un autre quart des observations se retrouve entre le premier quartile Q1 et le deuxième quartile Q2. Entre Q2 et Q3 on retrouve également un quart des valeurs, de même entre Q3 et la valeur maximum. Comme nous le verrons dans la partie « Paramètres de dispersion », les valeurs des quantiles sont utilisées pour définir des intervalles.

Notons que le deuxième quartile n'est autre que la médiane.

Exemple :

Soit une série des âges de n = 20 individus : 3, 5, 6, 7, 8, 11, 15, 20, 21, 22, 23, 23, 23, 30, 31, 32, 35, 36, 40, 45.

L'effectif de chaque quartile est donc de 5.

La valeur qui sépare le 1er groupe du 2ème est donc située entre 8 et 11.

Toute valeur comprise entre 8 et 11 peut être retenue comme premier quartile, toute valeur entre 22 et 23 comme deuxième quartile et toute valeur comprise entre 31 et 32 comme troisième quartile.

Les percentiles : Les percentiles définissent 100 groupes d'effectifs correspondants chacun à 1 % de l'effectif de l'échantillon. Le 50ème percentile est à la médiane.

Paramètres de dispersion

La moyenne ne suffit pas pour caractériser un ensemble de données.

Exemple :

La valeur moyenne de la série suivante : 1, 8, 9, 10, 11, 12, 19 est égale à 10.

La valeur moyenne de la série 8, 8, 9, 10, 11, 12, 12, est aussi égale à 10.

Dans le deuxième cas, la dispersion des mesures autour de la moyenne 10 est beaucoup moins importante que dans le premier cas.

Ces situations correspondent aux schémas de droite de la Figure 2.8.

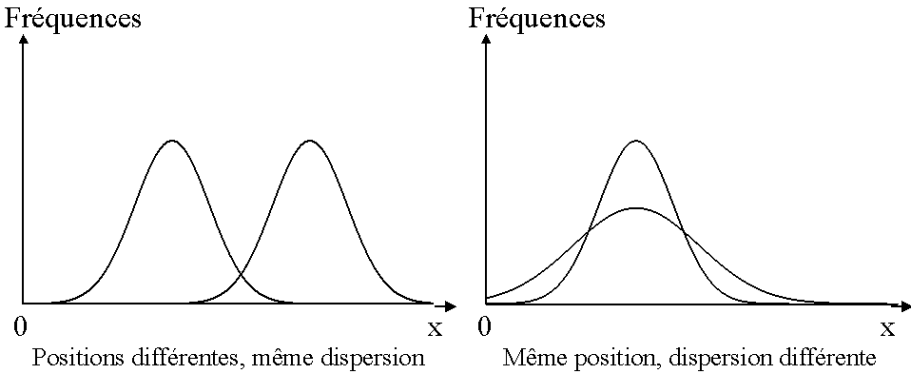

Figure 2.8 : Paramètres de dispersion et de position.

Variance et écart-type

Le paramètre le plus efficace pour rendre compte de la dispersion d'une série de mesures est la variance, ou sa racine carrée : l'écart-type.

Variance : La variance est définie comme la moyenne arithmétique des carrés des écarts à la moyenne de l'échantillon.

Cette définition répond à la formule :

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

où $\bar{x}$ est la moyenne de la série de mesures et *n* l'effectif.

Attention : Var(X) est la variance de l'échantillon, ce n'est ni la variance de la population dont est issu l'échantillon, ni l'estimation de la variance de la population (cf. chapitre « Estimation ponctuelle et intervalle de confiance »).

Ecart-type : afin de disposer d'un indice de dispersion qui s'exprime dans la même unité que la grandeur mesurée, on considère, en général, la racine carrée de la variance : l'écart-type.

$$\text{Ecart-type}(X) = \sqrt{Var(X)}$$

Exemple :

Calcul de la variance et de l'écart-type de la mesure des poids de 5 individus dans un échantillon de moyenne 72 kg (Tableau 2.8) :

Individus	***1***	***2***	***3***	***4***	***5***	***total***
x_i	*70*	*68,5*	*72,5*	*73*	*76*	
$(x_i - \bar{x})$	*-2*	*-3,5*	*+0,5*	*+1*	*+4*	***0***
$(x_i - \bar{x})^2$	*4*	*12,25*	*0,25*	*1*	*16*	*33,5*

Tableau 2.8 : Disposition des calculs pour la variance.

D'où la variance : 33,5 / 5 = 6,7 Kg^2 et l'écart-type : 2,59 Kg.

Propriété de la variance :

Soit la série $\{x_1, x_2, \ldots, x_n\}$ de moyenne $\bar{x}$, de variance Var(X) et deux constantes a et b.

Considérons la série $\{y_1, \ldots, y_n\}$ sachant que $y_i = ax_i + b$

soit $\{y_1 = ax_1 + b, y_2 = ax_2 + b, \ldots, y_n = ax_n + b\}$.

Le calcul montre que la variance de cette nouvelle série (que nous noterons Var(Y) est égal à :

$$\text{Var(Y)} = a^2\text{Var(X)}$$

$$\text{Ecart-type}(Y) = |a| \cdot \sqrt{Var(X)} = |a| \cdot \text{Ecart-type}(X)$$

Autres paramètres de dispersion.

Etendue : On définit l'étendue e d'une série de mesures comme la différence entre la plus grande et la plus petite valeur de la série :

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

L'**intervalle inter-quartile** représente 50 % des observations centrées en probabilité sur la médiane. Ses bornes sont Q1 et Q3. La largeur de cet intervalle, égale à Q3 - Q1, donne une idée de la dispersion des données : plus cette largeur est faible, plus les données sont groupées.

Remarque

Mode, médiane, quantiles peuvent être utilisés dans le cas de données ordinales.

Caractérisation des données quantitatives à deux dimensions

Introduction

Considérons des mesures X et Y effectuées sur un échantillon d'effectif *n*. Deux mesures sont effectuées sur le même individu « *i* » : x_i et y_i.

Exemple :

On peut mesurer chez n individus la concentration sanguine du potassium et du chlore. Nous obtenons pour chaque individu i le couple de mesures (x_i , y_i).

Représentation dans le plan

En reportant ces valeurs sur deux axes, on peut construire le point du plan correspondant aux coordonnées x_i et y_i. L'ensemble des points constitue « un nuage de points ». L'aspect de ce nuage est important à observer.

Caractérisation des couples (x_i , y_i)

Chacune des observations (x_i , y_i) est représentée par un point.

Le point ayant pour coordonnées $(\bar{x}\ ,\ \bar{y})$ représente le point « moyen » (Figure 2.9). Ce point, noté G, dont les coordonnées sont les moyennes des deux séries d'observations, s'appelle : « centre de gravité ».

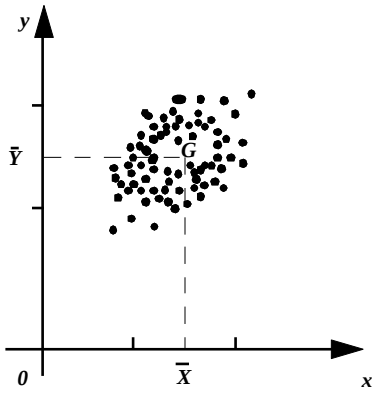

Figure 2.9 : Nuage de points.

Nous avions caractérisé la dispersion d'une donnée quantitative par la variance. D'une manière analogue, nous caractériserons la dispersion des points de coordonnées (x_i , y_i) par la moyenne du produit des écarts, pour chaque point, entre ses coordonnées et leurs valeurs moyennes. Nous définissons ainsi une nouvelle grandeur appelée **covariance**.

$$\text{Covar}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})$$

Coefficient de corrélation

Le coefficient de corrélation, noté *r*, est une grandeur qui reflète la dispersion des couples (x_i , y_i) en fonction de la dispersion observée pour chacune des deux séries de mesures X et Y.

$$r = \frac{\text{Covar}(X,Y)}{\sqrt{Var(X)\cdot Var(Y)}}$$

Remarque

Le coefficient de corrélation *r* est un nombre sans dimension. On montre qu'il est compris entre -1 et 1.

Ce qu'il faut savoir absolument

Représentation des différents types de données

Type de données	Présentation des données	Représentation graphique	Données à 2 dimensions
Qualitatif	Dénombrement par classes Fréquences absolues Fréquences relatives	Camembert	Tableau de contingence Regroupement libre
Ordinal	Dénombrement par classes ordonnées Fréquences absolues Fréquences relatives Fréquences cumulées	Diagramme en bâton	Tableau de contingence Regroupement des classes contiguës
Quantitatif	Tableau des mesures Paramètres de tendance centrale (ex : moyenne) Paramètres de dispersion (ex : variance, écart-type)	Histogramme	Nuage de points Covariance Coefficient de corrélation

Paramètres de tendance centrale d'une série d'observations

Type de données	Paramètre central	Calcul	Unité
Quantitatif	Moyenne	$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$	Celle des x_i
Quantitatif ou Ordinal	Médiane	Valeur(s) qui sépare(nt) l'ensemble des données ordonnées en deux sous-ensembles de même effectif	Celle des x_i
	Mode	Classe de plus grand effectif	Celle des x_i
	Quantiles	Valeur de la variable divisant l'ensemble ordonné des données en sous-ensembles d'effectifs égaux	Celle des x_i

Paramètres de dispersion d'une série d'observation

Type de données	Paramètre de dispersion	Calcul	Unité
Quantitatif	Variance	$Var(X) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}$	Celle des x_i au carré
	Ecart-type	$\sqrt{Var(X)}$	Celle des x_i
Quantitatif ou Ordinal	Intervalle inter-quartile	Différence entre les valeurs du troisième et du premier quartiles : Q3 – Q1	Celle des x_i

Chapitre 3 Notions de probabilité

Introduction

Le calcul de probabilité permet de modéliser des **phénomènes aléatoires**, c'est-à-dire des phénomènes pour lesquels les issues sont connues mais dont on ne peut en prédire la valeur car leur réalisation est incertaine.

> *Exemple :*
>
> *Concernant le sexe, les seules issues possibles, et connues, sont « masculin » et « féminin ». Cependant, pour un couple qui le désire, il n'est pas possible de prédire de manière certaine quel sera le sexe d'un enfant.*

L'observation des issues d'un phénomène sur des séries suffisamment grandes permet d'en déterminer leurs fréquences et par suite de connaître la loi qui le dirige (cf. section « Probabilités élémentaires »). Le calcul des probabilités permet de modéliser ces phénomènes aléatoires en attribuant à chacune de ses issues possibles une vraisemblance plus ou moins grande.

> *Exemple (suite) :*
>
> *On considère généralement que la probabilité d'être de sexe masculin est la même que la probabilité d'être de sexe féminin, soit 0,5.*

Evènements

Définitions

Ensemble fondamental

C'est l'ensemble des issues possibles d'un phénomène aléatoire, c'est-à-dire d'une expérience que l'on appelle généralement une **épreuve**.

> *Exemple :*
>
> *Le système ABO comporte 3 allèles : A, B et o. A et B sont co-dominants et dominent o. {AA, Ao, AB, BB, Bo, oo} est l'ensemble fondamental concernant le génotype d'un individu.*
>
> *L'épreuve est la détermination du groupe sanguin d'un individu tiré au sort.*

Evènements

Un événement correspond à un sous-ensemble d'un ensemble fondamental. On considère deux types d'évènements :

- Un **évènement élémentaire** correspond à une seule éventualité. L'ensemble des événements élémentaires constitue l'ensemble fondamental ;
- Un **évènement composé** correspond à la réunion (un regroupement) d'évènements élémentaires.

> *Par exemple, AA, AB et Ao sont des événements élémentaires alors que le groupe sanguin A (individus AA ou Ao) est un événement composé. L'ensemble fondamental est constitué des événements élémentaires AA, Ao, AB, BB, Bo et oo.*

Opérations sur les évènements

Il est possible de combiner des événements entre eux pour former de nouveaux événements. Soient E_1 et E_2 deux événements, on peut alors définir les opérations suivantes :

Union

$C = E_1 \cup E_2$: l'événement C est réalisé si et seulement si E_1 est réalisé OU E_2 est réalisé OU les deux sont réalisés.

Exemple :

BB ∪ Bo = Groupe sanguin B.

L'opération union permet donc de déterminer un événement composé.

Intersection

$C = E_1 \cap E_2$: l'événement C est réalisé si et seulement si E_1 ET E_2 sont réalisés tous les deux.

Exemple :

Allèle o ∩ homozygote = Groupe sanguin o.

Complémentarité

C = non E_1 : l'événement C est réalisé si et seulement si E_1 n'est pas réalisé.

Exemple :

Non allèle A est réalisé si le groupe sanguin est B ou o.

Remarque : On distingue 2 évènements particuliers :

- L'événement toujours réalisé ou **évènement certain** contenant tous les résultats possibles est noté Ω : $E_1 \cup$ non $E_1 = \Omega$.
- L'événement jamais réalisé ou **événement impossible** qui ne contient aucun des résultats possible est noté ∅ : $E_1 \cap$ non $E_1 = \varnothing$.

Evènements incompatibles ou disjoints

Deux événements sont incompatibles ou disjoints s'ils ne peuvent pas être réalisés simultanément, c'est-à-dire que leur intersection est l'événement impossible : $E_1 \cap E_2 = \varnothing$.

Exemple :

Groupe sanguin A et B sont disjoints car aucun individu n'est à la fois du groupe sanguin A et du groupe sanguin B.

Remarque 1 : deux événements élémentaires distincts sont disjoints.

Remarque 2 : un **ensemble d'événements** correspond à des événements élémentaires munis des opérations union, intersection et complémentarité.

Exemple :

Considérons un gène ayant 2 allèles A et a (A dominant) et la descendance de 2 hétérozygotes (Tableau 3.1) :

	A	***a***
A	*AA*	*Aa*
a	*aA*	*aa*

Tableau 3.1 : Evènements possibles pour un gène et deux allèles.

$\Omega = \{AA, Aa, aA, aa\}$

ξ *(ensemble d'événements) = AA, Aa, aA, aa, hétérozygote* $(Aa \cup aA)$*, homozygote* $(AA \cup aa)$*, phénotype A* $(AA \cup Aa \cup aA)$*, …*

Partition

Soit ξ un ensemble d'évènements et E_1, E_2, …, E_k des évènements appartenant à ξ, avec $E_i \neq \varnothing$ (i = 1, …, k). E_1, E_2, …, E_k forment une **partition** de Ω si et seulement si :

- Leur réunion est l'ensemble fondamental : $E_1 \cup E_2 \cup \ldots \cup E_k = \Omega$;
- Les évènements sont deux à deux disjoints : $\forall\ (i \neq j)$, $(E_i \cap E_j) = \varnothing$.

Exemple :

Les groupes sanguins A, B, AB et o forment une partition.

Probabilités

Probabilités élémentaires

Soit le phénomène aléatoire « détermination du groupe sanguin dans le système ABO » dont l'ensemble fondamental est Ω = {AA, Ao, BB, Bo, AB, oo}.

La répétition n fois de cette épreuve (c'est-à-dire qu'on recueille l'information

concernant le génotype sur *n* individus différents) permet de construire le Tableau 3.2 :

	AA	Ao	BB	Bo	AB	oo	Total
Fréquences absolues	n_1	n_2	n_3	n_4	n_5	n_6	n
Fréquences relatives	n_1/n	n_2/n	n_3/n	n_4/n	n_5/n	n_6/n	1

Tableau 3.2 : Fréquences observées et relatives lors de la répétition de *n* épreuves.

Propriétés :

- Si E est un événement quelconque et freq(E) sa fréquence relative, alors : $freq(E) \geq 0$.
- Si Ω est l'événement certain, alors : $freq(\Omega) = n/n = 1$.

Par ailleurs, considérons les événements :

E1 = groupe sanguin A = AA ∪ Ao

E2 = groupe sanguin B = BB ∪ Bo

E3 = allèle o = Ao ∪ Bo ∪ oo

dont les fréquences relatives sont :

$freq(E1) = (n_1 + n_2)/n$

$freq(E2) = (n_3 + n_4)/n$

$freq(E3) = (n_2 + n_4 + n_6)/n$

Considérons les événements E1 et E2 :

E1 ∪ E2 = AA ∪ Ao ∪ BB ∪ Bo

E1 ∩ E2 = ∅

$freq(E1 \cup E2) = (n_1 + n_2)/n + (n_3 + n_4)/n = freq(E1) + freq(E2)$

Considérons les événements E2 et E3 :

E2 ∪ E3 = BB ∪ Bo ∪ Ao ∪ Bo ∪ oo

E2 ∩ E3 = Bo ≠ ∅

$freq(E2 \cup E3) = (n_2 + n_3 + n_4 + n_6)/n$

$freq(E2 \cup E3) \neq (n_3 + n_4)/n + (n_2 + n_4 + n_6)/n = freq(E2) + freq(E3)$

Ceci se généralise pour donner la propriété suivante :

- Si E_1 et E_2 sont 2 **événements disjoints**, alors : $\mathbf{freq(E_1 \cup E_2) = freq(E_1) + freq(E_2)}$.

Si le phénomène aléatoire est observé un très grand nombre de fois ($n \to \infty$) nous admettrons que la fréquence relative d'un événement tend vers une limite que l'on appelle la **probabilité** de l'événement.

Exemple :

Fréquence relative du groupe sanguin AB = freq(AB) → 2/100

Fréquence relative du groupe sanguin A = freq(AA) + freq(Ao) → 48/100

On va alors définir une probabilité en lui attribuant les mêmes propriétés que celles définies précédemment pour les fréquences.

Définitions

Soit ξ un ensemble d'évènements défini sur l'ensemble fondamental Ω. On appelle probabilité une application P qui à tout événement E associe un nombre réel telle que :

1. $P(E) \geq 0, \forall E \in \xi$
2. $P(\Omega) = 1$
3. Si E_1 et E_2 sont des événements **disjoints** ($E_1 \cap E_2 = \varnothing$) alors $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

Remarques :

- La propriété 3 se généralise : si E_1, E_2, …, E_n sont des événements disjoints 2 à 2, alors $P(E_1 \cup E_2 \cup \ldots \cup E_n) = P(E_1) + P(E_2) + \ldots + P(E_n)$
- $0 \leq P(E) \leq 1$
- $P(E) = 1 - P(\text{non}E)$, puisque $E \cap \text{non}E = \varnothing$ et $E \cup \text{non}E = \Omega$
- $P(\varnothing) = 0$, puisque $\text{non}\Omega = \varnothing$
- $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

Exemple :

Soit une famille dans laquelle le groupe sanguin de la mère est A (de génotype Ao) et celui du père est AB. Quelle est la probabilité pour qu'un enfant soit du groupe sanguin A ? (on admettra que les événements élémentaires ont tous la même probabilité notée p).

Les événements élémentaires sont : Ω = {AA, AB, Ao, Bo}.

Calculons d'abord la probabilité d'un événement élémentaire :

$P(\Omega) = 1 = P(AA \cup AB \cup Ao \cup Bo)$

Les événements élémentaires étant disjoints, on a :

$P(\Omega) = 1 = P(AA) + P(AB) + P(Ao) + P(Bo)$

et puisqu'il y a 4 événements élémentaires de même probabilité, on a :

$P(\Omega) = 1 = 4.p$ et donc $p = 1/4$.

$P(\text{groupe A}) = P(AA \cup Ao)$

comme AA et Ao sont disjoints :

$P(\text{groupe A}) = P(AA) + P(Ao) = 2/4$.

Probabilités conditionnelles

Soient A et B deux événements quelconques. Dans certains cas, P(A) peut être

différente si l'événement B est déjà réalisé. Pour définir cette probabilité (probabilité de A sachant que B est réalisé), il faut se restreindre au sous-ensemble des résultats possibles de (A ∩ B) au sein des résultats possibles de B. On parle alors de **probabilité conditionnelle**.

Définition

Soit ξ un ensemble d'évènements définis sur l'ensemble fondamental Ω, P une probabilité sur ξ, A et B deux événements appartenant à ξ, avec P(B) > 0. La probabilité conditionnelle de A par rapport à B (probabilité de A sachant B) est définie par :

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

On obtient ainsi une relation très utilisée :

$$P(A \cap B) = P(A / B) \cdot P(B) = P(B / A) \cdot P(A)$$

Une probabilité conditionnelle est une probabilité, c'est-à-dire qu'elle satisfait aux 3 axiomes de la définition d'une probabilité.

Exemple :

Dans une population donnée, chaque individu est soit porteur d'une maladie M (P(M)=0,1), soit non porteur de M. Il existe un test permettant d'établir le diagnostic de M ; ce test est positif dans 13,5 % des cas. Par ailleurs, la probabilité d'être porteur de la maladie et d'avoir un test positif est de 4 %. Si le test est positif, qu'elle est la probabilité d'être porteur de M ?

$P(M \cap T^+) = 0,04$ et $P(T^+) = 0,135$. On recherche la probabilité $P(M / T^+)$.

Par définition de la probabilité conditionnelle, on a :

$P(M / T^+) = P(M \cap T^+) / P(T^+) = 0,04 / 0,135 \approx 0,3$.

Dans cet exemple, on voit bien que la probabilité d'être porteur de M est fortement modifiée quand l'information concernant le test est connue.

Indépendance en probabilité

Quels que soient A et B, A et B sont indépendants si et seulement si :

$$P(A \cap B) = P(A) \cdot P(B)$$

Remarques :

- Si A et B sont indépendants et P(A) > 0, P(B) > 0, alors :

P(A / B) = P(A ∩ B) / P(B) = P(A)·P(B) / P(B) = P(A).

De même, P(B / A) = P(B).

C'est-à-dire que deux événements A et B sont indépendants si la réalisation de B ne change pas la probabilité de A (autrement dit, la probabilité pour que A soit réalisé est

la même que B se soit produit ou non).

- Il ne faut pas confondre événements indépendants et événements disjoints. Si A et B sont disjoints on ne peut pas avoir $P(A \cap B) = P(A) \cdot P(B)$ si $P(A) > 0$ et $P(B) > 0$ puisque $P(A \cap B) = 0$. Donc, 2 événements disjoints et de probabilités non nulles ne sont jamais indépendants.

  Exemple : les lois de Mendel

  Le caractère couleur des pois est déterminé par un gène possédant deux allèles J (jaune) et v (vert), J domine v.

  Mendel croisait des pois jaunes homozygotes (JJ) et des pois verts (vv) et il obtenait à la première génération uniquement des pois jaunes (Tableau 3.3). Croisant ces pois de première génération il obtenait, à la deuxième génération, deux sortes de pois : des pois jaunes et des pois verts en proportion 3/4 jaunes et 1/4 verts (Tableau 3.4). En continuant à croiser par autofécondation, il constatait que :

  - *Les pois verts redonnent indéfiniment des pois verts.*
  - *Les pois jaunes se divisent en 2 catégories :*
    - *1/3 redonnent toujours exclusivement des pois jaunes ;*
    - *2/3 donnent dans leur descendance 3 jaunes pour 1 vert.*

  Cela correspond à un tirage au hasard d'un gène parmi les 2 de chacun des pois initiaux. Les 2 tirages étant indépendants, la probabilité de l'un quelconque des 4 évènements possibles est donc de (1/2)·(1/2) = 1/4.

Première génération		
	J (1/2)	**J** (1/2)
v (1/2)	Jv (1/4)	Jv (1/4)
v (1/2)	Jv (1/4)	Jv (1/4)

Tableau 3.3 : Première génération : hétérozygotes tous jaunes (les valeurs des probabilités associées sont données entre parenthèses).

La première génération est formée d'hétérozygotes Jv tous jaunes.

Dans la seconde génération, la probabilité du phénotype Jaune est : $P(Jaune) = P(JJ \cup Jv \cup vJ) = P(JJ) + P(Jv) + P(vJ) = 3/4$, puisque JJ, Jv et vJ sont disjoints.

Parmi ces pois jaunes 1/3 sont homozygotes :
$P(JJ/Jaune) = (1/4) / (3/4) = 1/3$.

Deuxième génération		
	J (1/2)	**v** (1/2)

J (1/2)	JJ (1/4)	vJ (1/4)
v (1/2)	Jv (1/4)	vv (1/4)

Tableau 3.4 : Deuxième génération (les probabilités associées sont données entre parenthèses).

Un autre caractère, graine lisse ou ridée, est aussi déterminé par un gène ayant 2 allèles L, r, L domine r.

Mendel croisait des pois jaunes à graines lisses doubles homozygotes (JJ, LL) avec des pois verts à graines ridées doubles homozygotes (vv, rr). Il obtenait à la première génération des pois jaunes à graines lisses. Croisant entre eux ces pois de première génération, il obtenait (Tableau 3.5) :

- *9/16 de pois jaunes à graines lisses ;*
- *1/16 de pois verts à graines ridées ;*
- *3/16 de pois jaunes à graines ridées ;*
- *3/16 de pois verts à graines lisses.*

Il en concluait que ces 2 gènes gouvernant des caractères différents ont des ségrégations indépendantes.

Ceci correspond pour chacun des pois parents à un tirage au hasard d'un gène pour chacun des caractères, ces 2 tirages étant indépendants.

Les pois de première génération issus de doubles homozygotes sont tous doubles hétérozygotes (Jv, Lr).

La probabilité de chacune des issues pour chaque pois parent à la deuxième génération est de (1/2)·(1/2) = 1/4 à cause de l'hypothèse de ségrégations indépendantes (2ème loi de Mendel).

		J (1/2)	**v** (1/2)	**L** (1/2	**r** (1/2)
		JL (1/4)	**Jr** (1/4)	**vL** (1/4)	**vr** (1/4)
J (1/2)	**JL** (1/4)	(JJ, LL) (1/16)	(JJ, Lr) (1/16)	(Jv, LL) (1/16)	(Jv, Lr) (1/16)
v (1/2)	**Jr** (1/4)	(JJ, rL) (1/16)	(JJ, rr) (1/16)	(Jv, rL) (1/16)	(Jv, rr) (1/16)
L (1/2	**vL** (1/4)	(vJ, LL) (1/16)	(vJ, Lr) (1/16)	(vv, LL) (1/16)	(vv, Lr) (1/16)
r (1/2)	**vr** (1/4)	(vJ, rL) (1/16)	(vJ, rr) (1/16)	(vv, rL) (1/16)	(vv, rr) (1/16)

Tableau 3.5 : Combinaison des allèles et probabilités associées.

La probabilité de chacune des issues finales est de 1/16 = (1/4)·(1/4) puisqu'on admet l'indépendance entre pois parents.

Par suite :

P(jaune à graine lisse) = P((JJ, LL) ∪ (JJ, Lr) ∪ (Jv, LL) ∪ (Jv, Lr) ∪ (JJ, rL) ∪ (Jv, rL) ∪ (vJ, LL) ∪ (vJ, Lr) ∪ (vJ, rL))

= P(JJ, LL) + … + P(vJ, rL)

= 1/16 + … + 1/16 = 9/16, puisque ces évènements sont élémentaires, donc disjoints.

NB : dans certains cas, si les deux gènes sont portés par le même chromosome, il peut y avoir « linkage » et les événements ne sont pas indépendants.

Théorème de Bayes

La formule $P(A \cap B) = P(A / B)\cdot P(B) = P(B / A)\cdot P(A)$, nous donne :

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

Le développement de cette formule va conduire à une formulation développée du théorème de Bayes. Ce développement requiert l'utilisation de la formule des probabilités totales :

Soient ξ un ensemble d'évènements, P une probabilité définie sur ξ. Considérons les événements A_1, A_2, …, A_k formant une partition et B appartenant à ξ. On définit la formule des **probabilités totales** par :

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + ... + P(B \cap A_k)$$

Exemple :

Dans une population donnée, chaque individu est soit porteur d'une maladie M, soit non porteur de M, et la proportion des M est 0,05. Par ailleurs, on sait qu'un test T est positif chez 80 % des porteurs de M et 10 % des non porteurs.

Question 1. Quelle est la probabilité qu'un individu pris au hasard soit positif pour T ?

On recherche $P(T^+)$. Sachant que $P(M) = 0,05$, $P(T^+/M) = 0,8$ et que $P(T^+ / nonM) = 0,1$ on en déduit que : $P(nonM) = 1 - P(M) = 0,95$

M et nonM forment une partition :

$M \cap nonM = \varnothing$

$M \cup nonM = \Omega$

Par application de la formule des probabilités totales :

$P(T^+) = P(T^+ \cap M) + P(T^+ \cap nonM)$

et par application de $P(A \cap B) = P(A / B).P(B)$, on a :

$P(T^+) = P(T^+ / M).P(M) + P(T^+ / nonM).P(nonM)$

$P(T^+) = 0,8.0,05 + 0,1.0,95 = 0,04 + 0,095 = 0,135$

Remarquons que l'apport des T^+ par les nonM est majoritaire (95 pour 135).

Question 2. Quelle est la probabilité des M parmi les T^+ ?

On recherche $P(M / T^+)$.

$P(M / T^+) = P(M \cap T^+) / P(T^+)$, par définition de la probabilité conditionnelle.

et comme $P(M \cap T^+) = P(T^+ / M).P(M)$ on a finalement :

$P(M / T^+) = (P(T^+ / M).P(M)) / P(T^+)$, qui correspond à la formule présentée au début du paragraphe « Théorème de Bayes ».

$P(M / T^+) = (0,8.0,05) / (0,135) = 40 / 135 \approx 0,3$.

De même : $P(nonM / T^+) = 95 / 135 \approx 0,7$.

Remarquons que bien que T soit beaucoup plus souvent positif chez les M que les nonM, la proportion des nonM parmi les T^+ est bien supérieure à celle des M.

Soient ξ un ensemble d'événements, P une probabilité définie sur ξ, A_1, A_2, …, A_k une partition et B appartenant à ξ. Le **théorème de Bayes** est :

$$P(A_i / B) = \frac{P(B / A_i) \cdot P(A_i)}{P(B / A_1) \cdot P(A_1) + ... + P(B / A_k) \cdot P(A_k)}$$

En effet, sachant que :

- $P(A / B) = P(B / A).P(A) / P(B)$
- $P(B) = P(B \cap A_1) + \ldots + P(B \cap A_k)$

Comme $P(B \cap A_i) = P(B / A_i) \cdot P(A_i)$, i = 1, …, k

on a $P(B) = P(B / A_1).P(A_1) + \ldots + P(B / A_k).P(A_k)$

et donc

$$P(A_i / B) = \frac{P(B / A_i) \cdot P(A_i)}{P(B)} = \frac{P(B / A_i) \cdot P(A_i)}{P(B / A_1) \cdot P(A_1) + ... + P(B / A_k) \cdot P(A_k)}$$

Classiquement, on appelle A_i les causes, $P(A_i)$ les probabilité « *a priori* », $P(A_i / B)$ les probabilités « *a posteriori* » et $P(B / A_i)$ les probabilités conditionnelles.

Exemple : groupes sanguins et filiation.

Le système ABO comporte 3 allèles A, B et o ; A et B sont co-dominants et dominent o.

La mère d'un enfant est de phénotype AB. Sachant que cet enfant est de phénotype B quelle est la probabilité que son père soit de phénotype A ?

Notons [B] l'événement « mère AB et enfant B ». Nous avons à déterminer les probabilités du phénotype du père sachant [B] : P(ph / [B]). Les phénotypes étant des réunions de génotypes, il est plus simple de calculer les probabilités du génotype du père sachant [B] puis de les regrouper.

Les probabilités de chaque génotype sont connues par ailleurs. Elles sont reportées dans la 3ème colonne du Tableau 3.6.

Puisque la mère est de phénotype AB et l'enfant de phénotype B, la mère a forcement donné un gène B. Donc pour que l'enfant soit B, le père a donné un gène B ou o. Par suite, P(ph / [B]) = P(ph / père donne B ou o).

Les probabilités de donner un gène B ou o sachant le génotype sont données dans la 4ème colonne du Tableau 3.6. On peut ainsi calculer pour chaque génotype la probabilité d'avoir le génotype en question et de donner B ou o sachant ce génotype (5ème colonne du Tableau 3.6).

Les divers génotypes formant une partition, nous pouvons appliquer le théorème de Bayes pour calculer P(génotype / [B]) (4ème colonne du Tableau 3.7).

Par suite, on a :

P(père A / [B]) = 0,195 / 0,705 = 0,277.

Phénotype	Génotype	P(génotype)	P(B ou o / génotype)	P(génotype) . P(B ou o / génotype)
A	**AA**	0,09	0	0
A	**Ao**	0,39	0,5	0,195
AB	**AB**	0,02	0,5	0,01
B	**BB**	0,02	1	0,02
B	**Bo**	0,06	1	0,06
O	**oo**	0,42	1	0,42
	Total			**0,705**

Tableau 3.6 : Probabilités et génotypes.

Phénotype	Génotype	P(génotype) . P([B] / génotype)	P(génotype / [B])	P(ph / [B])
A	**AA**	0	0 / 0,705	(0 + 0,195) / 0,705
A	**Ao**	0,195	0,195 / 0,705	
AB	**AB**	0,01	0,01 / 0,705	0,01 / 0,705
B	**BB**	0,02	0,02 / 0,705	(0,02 + 0,06) / 0,705
B	**Bo**	0,06	0,06 / 0,705	
O	**oo**	0,42	0,42 / 0,705	0,42 / 0,705
	Total	**0,705**	**1**	**1**

Tableau 3.7 : Probabilités et groupes ABO. Les probabilités P(ph/[B]) s'obtiennent par addition des P(génotype/[B]) puisque les divers génotypes sont disjoints.

Ce qu'il faut savoir absolument

Probabilités :

Il faut définir précisément les évènements auxquels on s'intéresse avant de faire un calcul de probabilité.

La probabilité d'un événement est toujours comprise entre 0 et 1.

Probabilités complémentaires	P(nonA) = 1 - P(A)
Propriété d'additivité	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ $P(A \cup B) = P(A) + P(B)$, si A et B sont disjoints
Probabilité conditionnelle	$P(A / B) = P(A \cap B) / P(B)$, si $P(B) \neq 0$
Indépendance en probabilité	A et B sont indépendants ssi $P(A \cap B) = P(A).P(B)$
Probabilités composées	$P(A \cap B) = P(A / B).\ P(B) = P(B / A).P(A)$, si $P(A) \neq 0$ et $P(B) \neq 0$.
Probabilités totales	$P(B) = P(B \cap A_1) + P(B \cap A_2) + \ldots + P(B \cap A_k)$, où A_1, A_2, …, A_k forment une partition et B appartient à l'ensemble des évènements
Théorème de Bayes	$P(A_i / B) = \frac{P(B / A_i) \cdot P(A_i)}{P(B / A_1) \cdot P(A_1) + \ldots + P(B / A_k) \cdot P(A_k)}$ $P(A_i / B) = \frac{P(B \cap A_i)}{P(B \cap A_1) + \ldots + P(B \cap A_k)}$

Chapitre 4 Variables aléatoires, lois de distribution

Exemple introductif

Un couple souhaitant avoir 2 enfants s'intéresse au nombre de garçons qu'il pourrait avoir. On admet que la naissance d'un garçon est aussi probable que celle d'une fille (P(G) = P(F) = 1/2) et que les naissances sont indépendantes. Le nombre de garçons dans cette fratrie ne peut pas être choisi par les parents ; il est régi par un phénomène aléatoire.

Notons X = nombre de garçons. Les valeurs possibles de X sont 0, 1 ou 2 avec des probabilités différentes. Le Tableau 4.1 donne la probabilité associée, c'est-à-dire celle de l'évènement correspondant, pour chaque valeur possible x de X. L'ensemble des valeurs possibles et leurs probabilités associées définissent la loi (ou distribution) de X.

Evènements	Valeur de X	Probabilité associée
F puis F	0	1/2.1/2 = 1/4
F puis G ou G puis F	1	1/2.1/2 + 1/2.1/2 = 1/2
G puis G	2	1/2.1/2 = 1/4

Tableau 4.1 : Probabilités associées à chaque valeurs possibles de X.

Ainsi, la probabilité de n'avoir aucun garçons est de 0,25, celle de ne pas avoir plus de 1 garçon est de 0,75 (P(FF ∪FG ∪GF) = 0,75 car les évènements élémentaires sont disjoints).

Il est possible de représenter graphiquement les probabilités associées aux valeurs possibles de X par un diagramme en bâton (Figure 4.1).

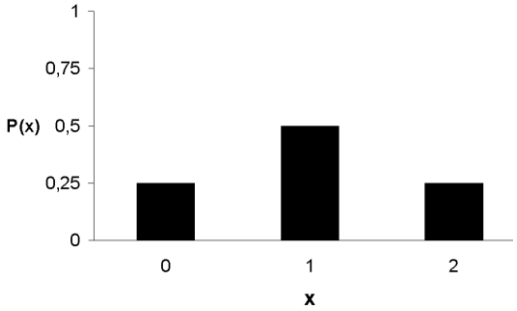

Figure 4.1 : Diagramme en bâtons de la distribution de probabilité d'avoir un garçon dans une

fratrie de 2 enfants.

Variables aléatoires discontinues ou discrètes

Définitions

Une **variable aléatoire** (v.a.) **discontinue** X prend différentes valeurs x avec des probabilités définies par sa distribution de probabilité p(x) (ou distribution de X).

La **distribution** de X est définie par l'ensemble des valeurs possibles x et de leurs probabilités associées $\{(x_1, P(X = x_1)), \ldots, (x_k, P(X = x_k))\}$.

Soient $x_1 \neq x_2 \neq \ldots \neq x_k$ les valeurs possibles prises par la v.a. X, on note $p_i = P(X = x_i)$ la probabilité pour que X prenne la valeur x_i, avec $p_1 + p_2 + \ldots + p_k = 1$.

Exemple (suite) :

Dans cet exemple, le nombre de garçon est une variable aléatoire pouvant prendre les valeurs x = 0, 1 ou 2.

Les probabilités associées sont $p_0 = P(X = 0) = 1/4$, $p_1 = P(X = 1) = 1/2$ et $p_2 = P(X = 2) = 1/4$.

La distribution du nombre de garçons dans une fratrie de 2 enfants correspond à l'ensemble $\{(x_0, p_0), (x_1, p_1), (x_2, p_2)\}$. Elle peut être représentée par un tableau (Tableau 4.1) ou par un diagramme en bâtons (Figure 4.1).

Propriété :

Nous admettrons que si X est une v.a. et si f est une fonction qui à tout nombre réel associe un nombre réel, alors f(X) est (en général) une v.a. de distribution $\{(f(x_i), p_i), i = 1, 2, \ldots, k\}$. Par exemple, $Y = aX + b$ (où a et b sont des constantes), X^2 sont des v.a.

Espérance mathématique ou moyenne d'une v.a. discrète

Soit X une v.a. de distribution $\{(x_i, p_i), i = 1, 2, \ldots, k\}$. L'**espérance mathématique** de X, ou **moyenne théorique**, est :

$$E(X) = x_1 p_1 + x_2 p_2 + \ldots + x_k p_k = \sum_{i=1}^{k} x_i p_i$$

Généralement l'espérance est notée μ.

Remarque :

Supposons qu'on répète un grand nombre N de fois une épreuve. On observe alors n_1 fois la valeur x_1, …, n_k fois la valeur x_k. La moyenne arithmétique de la variable mesurée serait (cf. chapitre « Statistique descriptive ») :

$$\bar{x} = \frac{n_1 x_1 + \ldots + n_k x_k}{N} = x_1 \frac{n_1}{N} + \ldots + x_k \frac{n_k}{N}$$

Or nous avons vu (cf. chapitre « Notion de probabilité ») que si un phénomène aléatoire est observé un très grand nombre de fois ($N \to \infty$) la fréquence relative de cet événement (n_i / N) tend vers la probabilité de réalisation de cet événement (p_i). Donc $\bar{x}$ tend vers μ.

Variance et écart-type d'une v.a. discrète

Soit X une v.a. de distribution $\{(x_i, p_i), i = 1, 2, \ldots, k\}$ et de moyenne μ. La **variance** de X, notée σ^2, est :

$$\sigma^2 = (x_1 - \mu)^2 p_1 + (x_2 - \mu)^2 p_2 + \ldots + (x_k - \mu)^2 p_k$$
$$= E\left((X - \mu)^2\right) = E\left((X - E(X))^2\right)$$

L'**écart-type** de X est égal à la racine carrée de la variance de X, soit σ. C'est une quantité positive.

Remarque : E(X), σ^2 et σ sont des constantes.

Propriétés concernant l'espérance mathématique et la variance d'une v.a. :

Soit X une v.a. de distribution $\{(x_i, p_i), i = 1, \ldots, k\}$, de moyenne μ et de variance σ^2.

Soit Y la v.a. définie par Y = aX + b (où a et b sont des constantes).

On démontre que (cf. infra) :

$E(Y) = a.E(X) + b = a.\mu + b$

$Var(Y) = a^2.Var(X) = a^2.\sigma^2$, et donc que l'écart-type de $Y = |a|.\sigma$

En effet : Y a pour distribution $\{(a.x_i + b, p_i), i = 1, 2, \ldots, k\}$.

$E(Y) = \Sigma(a.x_i + b)p_i = a\Sigma x_i p_i + b\Sigma p_i = a.\mu + b$

$Var(Y) = E[(Y - (a\mu + b))^2] = \Sigma(a.x_i - a\mu - b + b)^2 p_i = a^2\Sigma(x_i - \mu)^2 p_i = a^2.\sigma^2$

Un cas particulier sera utile par la suite : si on choisit $Y = (X - \mu) / \sigma$, on a alors E(Y) = 0 et Var(Y) = 1. On dit que Y est la v.a. **centrée réduite**.

Exemple :

Dans une population donnée, chaque individu est soit porteur d'une maladie M, soit non porteur de M, et la proportion des M est 0,05. Par ailleurs, on sait qu'un test T a, vis-à-vis de M, une sensibilité $Se = P(T^+ / M) = 0,8$ et une spécificité $Sp = P(T^- / nonM) = 0,9$.

On a créé une v.a. X indicatrice de l'erreur commise par le test, c'est-à-dire les faux positifs ou les faux négatifs (Tableau 4.2).

	Patient M	**Patient nonM**
T^+	0	+1
T^-	+1	0

Tableau 4.2 : Indicatrice de l'erreur du test.

Les probabilités associées à la v.a. X sont données dans le Tableau 4.3.

	Patient M	**Patient nonM**
T^+	Se.p = 0,8.0,05 = 0,04	(1 - Sp).(1 - p) = 0,1.0,95 = 0,095
T^-	(1 - Se).p = 0,2.0,05 = 0,01	Sp.(1 - p) = 0,9.0,95 = 0,855

Tableau 4.3 : Probabilités associées aux valeurs possibles de la v.a. X.

La distribution de la v.a. X est donc : {(0, (0,04 + 0,855)), (+1, (0,095 + 0,01))} = {(0, 0,895), (+1, 0,105)}.

La moyenne de X est : E(X) = 0.0,895 + 1.0,105 = 0,105.

E(X) peut être interprétée comme la moyenne du nombre d'erreurs sur un grand nombre de tests, ou bien comme la proportion d'erreurs, puisque X prend la valeur +1 si T conduit à faux positif et 0 sinon. Si on effectue ce test sur 1000 individus on peut s'attendre à commettre 0,105.1000 = 105 erreurs.

Variables aléatoires conjointes ou variable aléatoire à 2 dimensions

Exemple :

Considérons le sexe et le groupe sanguin d'un enfant à naître et de parents tous deux Ao. On définit 2 v.a. :

- *S : sexe, qui prendra la valeur 1 si c'est un « garçon » et 2 si c'est une « fille » ;*
- *Ph : phénotype du groupe sanguin, qui prendra la valeur 1 si c'est « A » et 0 sinon*

Le Tableau 4.4 représente les distributions conjointes de ces 2 v.a. où à chaque possibilité correspond un couple de valeur pour la v.a. « sexe » et pour la v.a « phénotype » ainsi que la probabilité associée. Les marginales du tableau correspondent aux distributions des v.a. « sexe » et « phénotype » considérées isolement.

		Phénotype		***Marginale***
		A	***Non A***	
Sexe (S)	***Garçon***	*[S = 1, Ph = 1] (0,375)*	*[S = 1, Ph = 0] (0,125)*	*S = 1 (0,5)*
	Fille	*[S = 2,Ph = 1] (0,375)*	*[S = 2, Ph = 0] (0,125)*	*S = 2 (0,5)*

Marginale		*Ph = 1 (0,75)*	*Ph = 0 (0,25)*	

Tableau 4.4 : Variables aléatoires conjointes **indépendantes** (les probabilités associées sont données entre parenthèses).

La distribution de la v.a. « sexe » est {(1, 0,5), (2, 0,5)}.

La distribution de la v.a. « phénotype » est {(1, 0,75), (0, 0,25)}.

La distribution du couple de v.a. « sexe » et « phénotype » est : {((1 , 1), 0,375), ((2 , 1), 0,375), ((1 , 0), 0,125), ((2 , 0), 0,125)}.

Variables aléatoires indépendantes

Soient X, une v.a. de distribution $\{(x_i, p_i), i = 1, 2, \ldots, n_x\}$, et Y, une autre v.a. de distribution $\{(y_j, p_j), j = 1, 2, \ldots, n_y\}$, c'est-à-dire que X et Y sont des v.a. conjointes.

La distribution du couple de v.a. X et Y est notée : $\{(x_i, y_j), r_{ij} = P(X = x_i \text{ ET } Y = y_j), i = 1, \ldots, n_x \text{ et } j = 1, \ldots, n_y\}$.

X et Y sont indépendantes si et seulement si :

$r_{ij} = P(X = x_i \text{ ET } Y = y_j) = P(X = x_i).P(Y = y_j) = p_i.q_j$, pour tout i et tout j.

Dans l'exemple précédent, le sexe est le phénotype du groupe sanguin sont indépendants puisque le contenu r_{ij} de chaque case du Tableau 4.4 est le produit des probabilités marginales correspondantes (P[S = 1, Ph = 1] = 0,375 = P(G).P(A) = 0,5.0,75).

Propriétés

- Soient $X_1, X_2, \ldots, X_n$ des v.a. conjointes, alors :

  $E(X_1 + X_2 + \ldots + X_n) = E(X_1) + E(X_2) + \ldots + E(X_n)$

- Si $X_1, X_2, \ldots, X_n$ sont indépendantes 2 à 2, alors :

  $Var(X_1 + X_2 + \ldots + X_n) = Var(X_1) + Var(X_2) + \ldots + Var(X_n)$

  $E(X_1.X_2) = E(X_1).E(X_2)$

- En particulier, si X_1 et X_2 sont indépendantes alors,

  $Var(X_1 - X_2) = Var(X_1) + (-1)^2 Var(X_2) = Var(X_1) + Var(X_2)$

  Exemple :

  L'hémophilie est une maladie héréditaire récessive liée au chromosome X.

  Considérons un couple dans lequel la mère est porteuse saine de l'anomalie chromosomique et le père est sain et désirant avoir un enfant. Dans cette situation, la prévalence de l'hémophilie exprimée est alors de 0,25.

  Définissons une première v.a. « sexe » qui prendra la valeur 1 si c'est un « garçon » et 2 si c'est une « fille » et une deuxième v.a. « hémophilie » qui prendra la valeur 1 en « présence » d'hémophilie exprimée et 0 sinon. Le Tableau 4.5 représente les distributions conjointes de ces 2 v.a.

		Hémophilie		*Marginale*
		Présence	***Absence***	
Sexe	***Garçon***	*[S = 1, Hémo. = 1] (0,25)*	*[S = 1, Hémo. = 0] (0,25)*	*Sexe = 1 (0,5)*
	Fille	*[S = 2, Hémo. = 1] (0)*	*[S = 2, Hémo. = 0] (0,5)*	*Sexe = 2 (0,5)*
Marginale		*Hémophilie = 1 (0,25)*	*Hémophilie = 0 (0,75)*	

Tableau 4.5 : Variables aléatoires conjointes **non indépendantes** (les probabilités associées sont données entre parenthèses).

Dans cet exemple, le sexe et l'hémophilie ne sont pas indépendants puisque P[S = 1, Hémo. = 1] = 0,25 ≠ P[S = 1].P[Hémo. = 1] = 0,5.0,25.

Deux v.a qui ne sont pas indépendantes sont liées entre elles. On définit alors une **distribution conditionnelle** :

Soient X et Y des v.a. conjointes de distribution $\{(x_i, y_j), r_{ij} = P(X = x_i \text{ ET } Y = y_j), i = 1, \ldots, n_x \text{ et } j = 1, \ldots, n_y\}$.

La distribution de Y lorsque X = xi (notée Y / X = xi) est définie par :

$\{(x_i, y_j), r_{j/i} = P(X = x_i \text{ ET } Y = y_j) / P(X = x_i), j = 1, \ldots, n_y\}$.

Covariance, coefficient de corrélation

Soient X et Y des v.a conjointes de distribution $\{(x_i, y_j), r_{ij} = P(X = x_i \text{ ET } Y = y_j), i = 1, \ldots, n_x \text{ et } j = 1, \ldots, n_y\}$. Soient μ_X la moyenne et σ_X^2 la variance de X, μ_Y la moyenne et σ_Y^2 la variance de Y.

La covariance de X et Y est :

$$\text{Covar}(X,Y) = E((X-\mu_X).(Y-\mu_Y)) = \sum_{i=1}^{n_X}\sum_{j=1}^{n_Y}(x_i-\mu_X).(y_j-\mu_Y).r_{ij}$$

Le coefficient de corrélation est :

$$\rho_{XY} = \frac{\text{Covar}(X,Y)}{\sigma_X.\sigma_Y}$$

Propriétés

- ρ_{XY} est un nombre sans dimensions tel que $-1 \leq \rho_{XY} \leq +1$
- Si X et Y sont indépendantes alors $\rho_{XY} = 0$ (la réciproque est en général fausse)

- On peut démontrer que lorsque Y = aX + b, alors $\rho_{XY} = \pm 1$ avec le signe de a ($a \neq 0$)

Exemple sur le sexe et le phénotype du groupe sanguin (suite)

Utilisons une v.a. S indicatrice du sexe telle que s = 1 pour les garçons et s = 0 pour les filles ainsi qu'une v.a. Ph indicatrice du phénotype telle que ph = 1 pour les phénotype A et ph = 0 les phénotype non A.

La distribution du couple de v.a. S et Ph est notée : {((1 1), 0,375), ((0 1), 0,375), ((1 0), 0,125), ((0 0), 0,125)}.

On a pour la v.a. S :

$E(S) = 1.0,5 + 0.0,5 = 0,5$

$Var(S) = (1 - 0,5)^2.0,5 + (0 - 0,5)^2.0,5 = (0,5)^2$

$\sigma_S = 0,5$

et pour la v.a. Ph :

$E(Ph) = 1.0,75 + 0.0,25 = 0,75$

$Var(Ph) = (1 - 0,75)^2.0,75 + (0 - 0,75)^2.0,25 = 0,19$

$\sigma_{Ph} = 0,43$

$Covar(S, Ph) = (1 - 0,5).(1 - 0,75).0,375 + (1 - 0,5).(0 - 0,75).0,125$
$+ (0 - 0,5).(1 - 0,75).0,375 + (0 - 0,5).(0 - 0,75).0,125 = 0$

$\rho_{S,Ph} = 0$

Exemple sur le sexe et l'hémophilie (suite)

Utilisons à nouveau la v.a. S indicatrice du sexe telle que s = 1 pour les garçons et s = 0 pour les filles et une v.a. H indicatrice de l'hémophilie telle que h = 1 pour hémophilie = oui et h = 0 pour hémophilie = non.

La distribution du couple de v.a. S et H est notée : {((1 1), 0,25), ((0 1), 0), ((1 0), 0,25), ((0 0), 0,5)}.

On a pour la v.a. S :

$E(S) = 0,5$; $Var(S) = (0,5)^2$; $\sigma_S = 0,5$

et pour la v.a. H :

$E(H) = 1.0,25 + 0.0,75 = 0,25$

$Var(H) = (1 - 0,25)^2.0,25 + (0 - 0,25)^2.0,75 = 0,19$

$\sigma_H = 0,43$

$Covar(S, H) = (1 - 0,5).(1 - 0,25).0,25 + (1 - 0,5).(0 - 0,25).0,25$
$+ (0 - 0,5).(1 - 0,25).0 + (0 - 0,5).(0 - 0,25).0,5 = 0,13$

$\rho_{S,H} = 0,13/(0,5.0,43) = 0,58$

Variables aléatoires continues

Les v.a. étudiées précédemment étaient définies sur un ensemble d'évènements en nombre fini (nombre de garçons dans une fratrie de 2 enfants, phénotype du groupe sanguin A ou non A en fonction du sexe, …). On peut également définir des v.a. lorsque l'ensemble d'évènements est infini, c'est-à-dire lorsque l'on s'intéresse à une variable continue.

La généralisation du cas des v.a. discontinues au cas de v.a. continues peut être abordée de manière intuitive à partir de l'histogramme du polygone des fréquences (cf. chapitre « Statistique descriptive ») : lorsque la taille de l'échantillon devient infinie et la largeur des classes tend vers 0, alors la limite du polygone des fréquences tend vers la **densité de probabilité** d'une v.a. continue (Figure 4.2).

Une densité de probabilité f(X) est positive ou nulle.

La probabilité pour qu'une réalisation au hasard de la v.a. soit comprise entre deux valeurs x_1 et x_2 correspond à la surface comprise entre la courbe de densité et l'axe des X limité par les 2 verticales passant par x_1 et x_2 (Figure 4.2).

Remarque :

- Il en résulte que $P(x_1 \leq X \leq x_1) = 0$, c'est-à-dire que $P(X = x_1) = 0$.
- La surface délimitée par la courbe de densité et l'axe des X sans bornes vaut 1 (la somme de toutes les probabilités élémentaires vaut 1 : $\int_{-\infty}^{+\infty} f(x)dx = 1$.

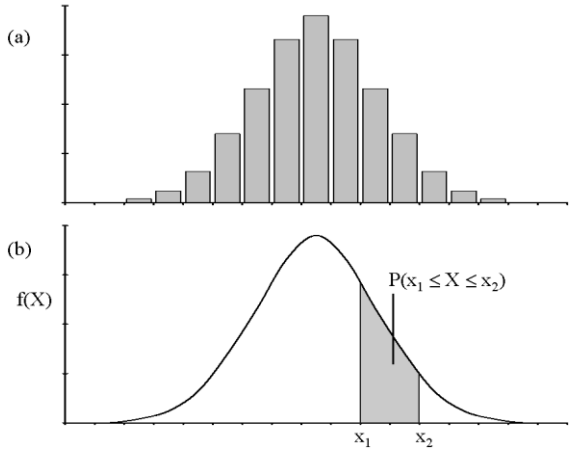

Figure 4.2 : Approche d'une densité de probabilité à partir d'un histogramme. (a) Histogramme des fréquences. L'augmentation de la taille de l'échantillon et la réduction de la largeur des classes tend vers (b) la densité de probabilité f(X) d'une v.a continue.

Le passage du cas discontinu au cas continu revient à considérer la probabilité de trouver une variable X dans un intervalle donné ce qui transforme les sommes Σ en intégrales ∫ et les p_i en f(x)dx.

L'espérance mathématique est donc :

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)dx$$

La moyenne de X a la même signification que dans le cas discontinu : c'est la limite de la moyenne arithmétique d'un échantillon lorsque la taille tend vers l'infini.

La variance est toujours définie par :

$$\sigma^2 = E\left((X - E(X))^2\right)$$

Lois de distribution

Nous présentons ici trois lois de distribution utilisées dans la suite de ce cours.

Loi Normale

La distribution Normale, ou de Laplace Gauss, ne dépend que de 2 paramètres : la moyenne, μ, et l'écart-type, σ. Nous noterons N(μ, σ) une v.a. Normale de moyenne μ et d'écart-type σ :

Propriétés de la loi Normale

- f(x) est totalement déterminée par sa moyenne et son écart-type ;
- La fonction de densité est (Figure 4.3) :
  - continue ;
  - symétrique par rapport à la moyenne μ ;
  - passe par un maximum pour x = μ (c'est-à-dire que le mode = μ) ;
  - a une médiane égale à μ ;
- Si X_1, X_2, …, X_n sont Normales et indépendantes alors $Y = X_1 + X_2 + \ldots + X_n$ est Normale ;
- Si X est N(μ, σ) alors Y = aX + b (a et b sont des constantes) est N(aμ + b, | a |.σ). Cette propriété permet d'établir un cas particulièrement utile par la suite en définissant une nouvelle v.a. Z telle que Z = (X - μ) / σ. Dans ce cas la loi de distribution de Z est N(0, 1), appelée **loi Normale centrée réduite** (la distribution est centrée sur 0 avec un écart-type égal à 1).

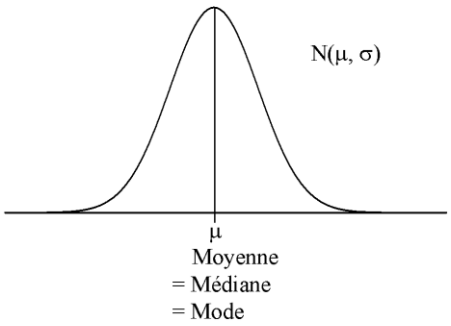

Figure 4.3 : Loi Normale de moyenne μ et d'écart-type σ.

Table de la loi Normale (cf. annexe)

Cette table concerne la **loi N(0, 1)**. Elle permet de déterminer la probabilité α pour

que Z dépasse une certaine valeur N_α. Autrement dit, elle donne pour certaines valeurs de α la valeur N_α telle que :

$$1-\alpha = \text{Proba}\left(-N_\alpha \le Z \le +N_\alpha\right) = \text{Proba}\left(|Z| \le N_\alpha\right)$$

soit

$$\alpha = \text{Proba}\left(\left(Z \le -N_\alpha\right) \text{ ou } \left(Z \ge +N_\alpha\right)\right) = \text{Proba}\left(|Z| \ge N_\alpha\right)$$

Toute v.a. Normale peut être rendue Normale centrée réduite ce qui autorise l'utilisation de la table. En effet, si Y est N(μ, σ) alors Z = (Y - μ,) / σ est N(0, 1) et donc (Figure 4.4) :

$$1-\alpha = \text{Proba}\left(\mu - \left(N_\alpha \cdot \sigma\right) \le Y \le \mu + \left(N_\alpha \cdot \sigma\right)\right)$$

et

$$\alpha = \text{Proba}\left(\left(Y \le \mu - \left(N_\alpha \cdot \sigma\right)\right) \text{ ou } \left(Y \ge \mu + \left(N_\alpha \cdot \sigma\right)\right)\right)$$

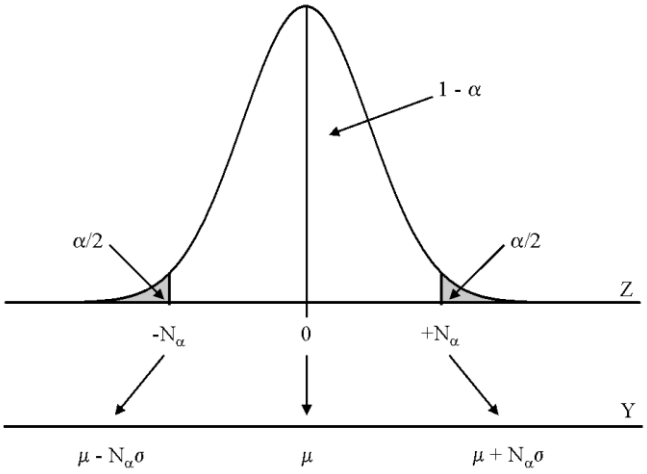

Figure 4.4 : Variable Normale et variable Normale centrée réduite.

Par ailleurs, les propriétés de la loi Normale impliquent que si :

$$\alpha_1 < \alpha_2 \text{ alors } N\alpha_1 > N\alpha_2$$

Certaines valeurs sont souvent utilisées dans la table de la loi Normale centrée réduite :

$N_{0,10} = 1,645$ $N_{0,05} = 1,96$ $N_{0,01} = 2,576$

Loi de Student

La loi de Student dépend d'un seul paramètre : son nombre de degré de liberté (ddl).

Le **nombre de degré de liberté** ν est une quantité exprimant le nombre de données indépendantes. Il n'y a pas une distribution de Student mais une famille de distributions de Student, une par ddl (Figure 4.5).

Propriétés de la loi de Student à ν ddl (Figure 4.5)

- Elle est symétrique par rapport à 0 ;
- Elle passe par un maximum pour 0 (le mode = 0) ;
- Elle est d'autant plus aplatie que ν est petit ;
- Elle tend vers la loi N(0, 1) lorsque ν tend vers l'infini.

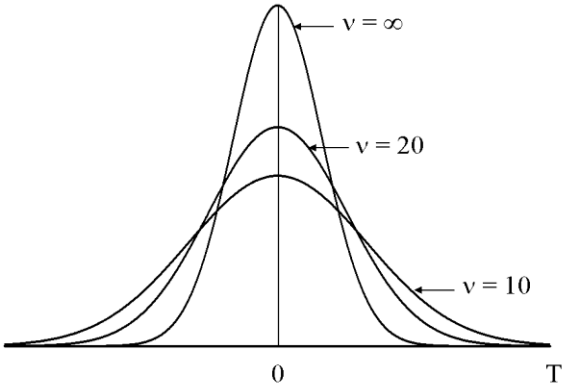

Figure 4.5 : Loi de Student pour certains degrés de liberté.

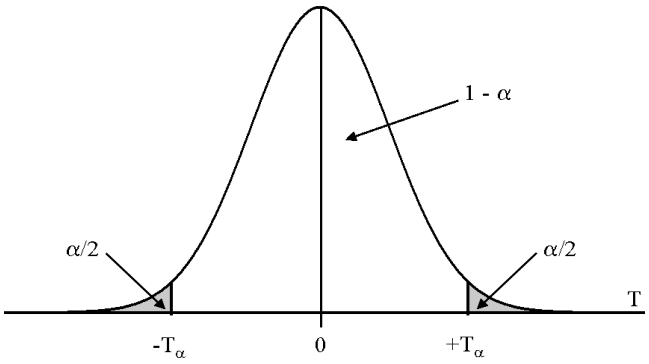

Figure 4.6 : Loi de Student à ν degrés de liberté et bornes au risque α.

Table de la loi de Student (cf. annexe)

Elle donne pour certaines valeurs de α la valeur $T_{\alpha,\nu}$ telle que (Figure 4.6) :

$$\alpha = \text{Proba}\left(\left(T \le -T_{\alpha,\nu}\right) \text{ ou } \left(T \ge +T_{\alpha,\nu}\right)\right) = \text{Proba}\left(|T| \ge T_{\alpha,\nu}\right)$$

Exemple :
si α = 5 % et ν = *10 alors* $T_{0,05\ ;\ 10} = 2{,}228$
si α = 10 % et ν = *8 alors* $T_{0,10\ ;\ 8} = 1{,}860$

Pour ν supérieur ou égal à 30, $T_{\alpha,\nu}$ est arrondi à N_α.

Loi du Chi-deux (χ^2)

La loi du Chi-deux ne dépend également que de son nombre de degrés de liberté. Il y a donc une famille de distribution de probabilité du χ^2 (Figure 4.7).

Propriétés de la loi du Chi-deux à ν ddl (Figure 4.7**)**

- La loi du χ^2 est asymétrique pour des « petites » valeurs de ν ;

Table de la loi du Chi-deux (cf. annexe)

Elle donne la probabilité α que χ^2 soit supérieur ou égal à une valeur donnée (Proba($\chi^2 \geq \chi^2_{\alpha,\nu}$)) pour chaque degré de liberté.

Par exemple (cf. Figure 4.7),
si $\alpha = 5$ % et $\nu = 1$ *alors* $\chi^2_{0,05,1} = 3{,}84$
si $\alpha = 5$ % et $\nu = 5$ *alors* $\chi^2_{0,05,5} = 11{,}07$

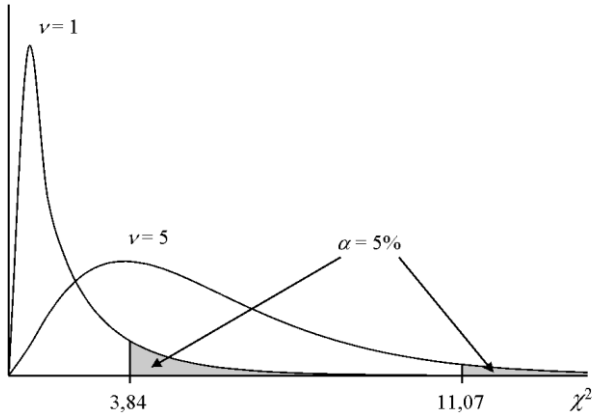

Figure 4.7 : La loi du χ^2.

Ce qu'il faut savoir absolument

Variable aléatoire discrète	
Espérance mathématique	$E(X) = \mu = x_1 p_1 + ... + x_k p_k = \sum_{i=1}^{k} x_i p_i$
Variance	$Var(X) = \sigma^2 = (x_1 - \mu)^2 p_1 + ... + (x_k - \mu)^2 p_k$ $= E\left((X-\mu)^2\right) = E\left((X - E(X))^2\right)$

Variable aléatoire continue	
Espérance mathématique	$E(X) = \int_{-\infty}^{+\infty} x \cdot f(x) dx$
Variance	$Var(X) = \sigma^2 = E\left((X - E(X))^2\right)$

Propriétés

- Espérance (somme de v.a.) = somme des espérances
- Variance (somme de v.a. indépendantes) = somme des variances

Variables aléatoires conjointes

Soient X et Y des v.a. conjointes de distribution $\{(x_i, y_j), r_{ij} = P(X = x_i \text{ ET } Y = y_i), i = 1, \ldots, n_x \text{ et } j = 1, \ldots, n_y\}$ où X a pour moyenne μ_X et pour variance σ_X^2 et Y a pour moyenne μ_Y et pour variance σ_Y^2 :

Covariance	$\text{Covar}(X,Y) = E((X-\mu_X).(Y-\mu_Y)) = \sum_{i=1}^{n_X}\sum_{j=1}^{n_Y}(x_i-\mu_X).(y_j-\mu_Y).r_{ij}$
Coefficient de corrélation	$\rho_{XY} = \frac{\text{Covar}(X,Y)}{\sigma_X.\sigma_Y}$

Propriétés du coefficient de corrélation :

- ρ_{XY} est un nombre sans dimensions tel que $-1 \leq \rho_{XY} \leq +1$
- Si X et Y sont indépendantes alors $\rho_{XY} = 0$ (la réciproque est en général fausse)

Loi de distribution Normale :

Pour utiliser la table de la loi Normale, il faut transformer la variable X qui suit une loi Normale $N(\mu, \sigma)$ en une variable centrée réduite N(0, 1).

Chapitre 5 Estimation ponctuelle et intervalle de confiance

Introduction

Il est peu fréquent d'étudier un caractère sur l'ensemble de la **population.** On travaille donc sur un **échantillon** extrait de la population.

Etant donné un résultat obtenu à partir d'un échantillon, que peut-on déduire sur la population dont il est issu, quelle inférence statistique peut-on faire ?[2] Par exemple, si le paramètre étudié est la moyenne, quelle est la valeur que l'on doit admettre pour la population à partir de la valeur calculée sur l'échantillon (Figure 5.1) ? Nous sommes ici dans un problème d'**estimation ponctuelle**.

Il n'y a pas forcement une estimation ponctuelle unique et il existe un ensemble de valeurs possibles, compatibles avec les observations, dans lequel on peut penser qu'est réellement située la valeur du paramètre de la population ; on parle alors d'**intervalle de confiance**. Il importe alors de fournir l'estimation la plus « vraisemblable » et de connaître la « précision » de cette estimation.

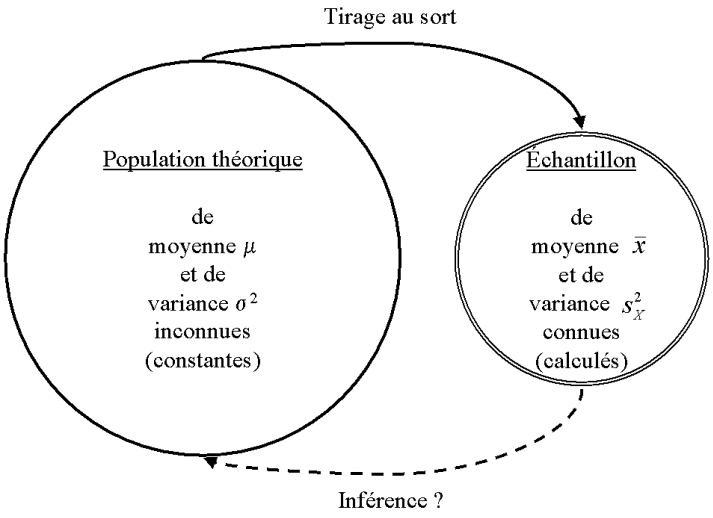

Figure 5.1 : Estimation des paramètres d'une population à partir d'un échantillon tiré au sort et inférence.

[2] La vraie valeur de la caractéristique dans la population est inconnue. On cherche à l'approcher à partir de calculs réalisés sur un échantillon.

Échantillon, estimateur et estimation

Un échantillon est une partie de la population cible. Un « bon » échantillon est un **échantillon représentatif** de la population cible, c'est-à-dire que les proportions des caractéristiques des éléments de l'échantillon sont très proches de celles de la population.

Une méthode d'**échantillonnage** par **tirage au sort** offre le maximum de garanties pour obtenir un échantillon représentatif. Ainsi, chaque élément de la population a la même probabilité de faire partie de l'échantillon.

On mesure ensuite sur chaque éléments constituant l'échantillon la caractéristique faisant l'objet de l'étude. On considère que la caractéristique est une variable aléatoire pour laquelle on veut connaître la distribution dans la population (par exemple, à travers sa moyenne et sa variance). Pour cela, on peut réaliser une **estimation ponctuelle** qui revient à attribuer une valeur, l'**estimation**, au paramètre de la population à partir des données provenant de l'échantillon. On est ainsi amené à construire un **estimateur** qui est une fonction qui associe l'estimation à l'échantillon.

> *Exemple :*
>
> *On s'intéresse au nombre de caries dentaires chez les enfants scolarisés dans le primaire à Marseille. Le rectorat de l'académie possède une liste exhaustive des enfants scolarisés dans le primaire à Marseille.*
>
> *On réalise, à partir de cette liste, un tirage au sort afin de constituer un échantillon représentatif de 300 enfants. Un examen dentaire est réalisé pour chacun d'eux afin de compter le nombre de caries par enfants.*
>
> *A partir de cet échantillon on obtient une estimation du nombre moyen de caries dentaires par enfants qui est une valeur approchée du nombre moyen de caries dentaires chez les enfants scolarisés dans le primaire à Marseille.*
>
> *Cette estimation est obtenue par la formule classique du calcul d'une moyenne. On montre que ce calcul fournit un « bon » estimateur (cette notion est définie dans le paragraphe suivant).*

Notons déjà que, d'un échantillon à un autre, l'estimateur est le même mais on peut avoir des estimations différentes (cela est dû aux fluctuations d'échantillonnage, c'est-à-dire au hasard du tirage).

Propriétés d'un « bon » estimateur

Biais

Un bon estimateur doit être sans biais. Soit θ un paramètre quelconque de la population et U un estimateur de θ :

- U est un **estimateur sans biais** de θ si $E(U) = \theta$
- U est un **estimateur biaisé** de θ si $E(U) \neq \theta$; le biais vaut : $E(U) - \theta$

Ces notions sont illustrées sur la Figure 5.2.

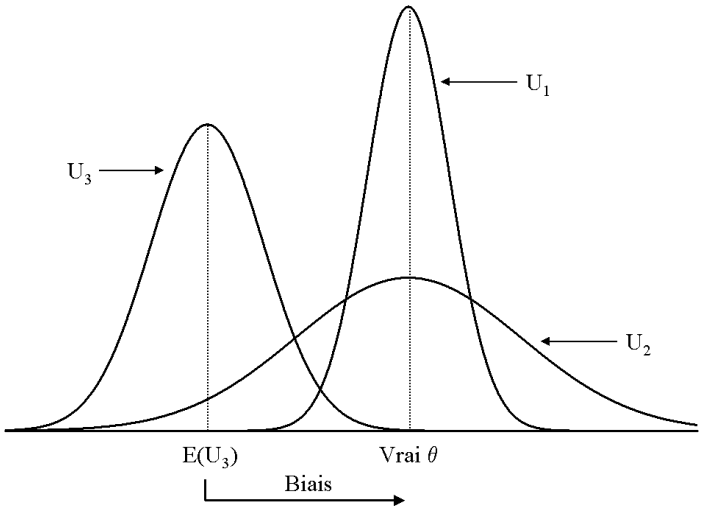

Figure 5.2 : Biais et variance pour 3 estimateurs d'un paramètre θ : U_1 et U_2 sont 2 estimateurs sans biais avec $Var(U_1) < Var(U_2)$; U_3 est un estimateur biaisé.

Variance

Un bon estimateur doit avoir une faible variance.

On dira d'un estimateur qu'il est **convergent** lorsqu'il est sans biais et que sa variance tend vers 0 quand l'effectif de l'échantillon observé tend vers l'infini.

Si deux estimateurs sont sans biais, le plus efficace est celui dont la variance est la plus petite puisque ses valeurs sont en moyenne plus proches du paramètre estimé (cf. Figure 5.2).

Estimation ponctuelle

Estimation de la moyenne et de la variance d'une population

Estimation de la moyenne d'une population

Soient μ et σ^2 la moyenne et la variance (inconnues), obtenues à partir d'un échantillon pris au hasard, d'une v.a. que l'on cherche à estimer. Supposons que l'on effectue p échantillonnages (tirages au sort), tous d'effectif n, dans cette population et que l'on obtienne les résultats suivants :

$x_1, x_2, \ldots, x_n$: premier échantillon d'effectif n

$y_1, y_2, \ldots, y_n$: deuxième échantillon d'effectif n

…

$z_1, z_2, \ldots, z_n$: $p^{ième}$ échantillon d'effectif n

On peut, pour chacun des échantillons, calculer leur moyenne :

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

…

$$\bar{z} = \frac{\sum_{i=1}^{n} z_i}{n}$$

Il est alors naturel de penser :

- Que chacune des moyennes $\bar{x}$, $\bar{y}$, …, $\bar{z}$ est une estimation de la moyenne de la population ;
- Qu'il n'est pas étonnant, par ailleurs, de trouver $\bar{x} \neq \bar{y} \neq ... \neq \bar{z}$.

Exemple :

On s'intéresse à la taille du nourrisson de sexe masculin, normal, à l'âge de trois mois. Soient μ et σ^2 la moyenne et la variance de cette v.a., paramètres que l'on cherche à estimer. On effectue p échantillonnages, tous d'effectifs $n = 8$, dans cette population. Les résultats obtenus sont donnés dans le Tableau 5.1. Chacune des moyennes est une estimation de la moyenne de la population et chaque moyenne est différente d'un échantillon à l'autre.

Echantillon 1	***Echantillon 2***	***…***	***Echantillon p***
62,8	*58,6*	*…*	*55,4*
54,4	*58,6*	*…*	*67,5*
56,9	*58,6*	*…*	*59,3*
62,6	*58,3*	*…*	*61,1*
58,5	*58,5*	*…*	*65,2*
60,5	*64,4*	*…*	*58,3*
66,3	*59,2*	*…*	*63,0*
64,0	*57,6*	*…*	*59,6*

Moyenne	***60,8***	***≠***	***59,2***	***≠***	***…***	***≠***	***61,2***

Tableau 5.1 : Valeurs moyennes de la taille (en cm) de p échantillons d'effectif n = 8 tirés au sort dans une population de nourrissons de sexe masculin, normaux, âgés de 3 mois.

L'ensemble des valeurs $\bar{x}$, $\bar{y}$, …, $\bar{z}$ constitue des observations d'une variable aléatoire $\bar{X}$ dont la loi de distribution est appelée **loi de distribution d'échantillonnage de la moyenne**. On démontre que cette loi a pour

- moyenne μ
- écart-type $\sigma_m = \sigma/\sqrt{n}$

La v.a. $\bar{X}$ est l'**estimateur** de μ.

C'est un estimateur **sans biais** puisque $\mathrm{E}(\bar{X}) = \mu$.

C'est un estimateur **convergent** puisqu'il est sans biais et que $Var(\bar{X}) = \sigma^2/n$ tend vers 0 lorsque n tend vers l'infini.

L'observation $\bar{x}$ est une bonne estimation de la moyenne μ de la population.

Exemple (suite) :

Considérons une population de N = 200 nourrissons de sexe masculin, normaux, à l'âge de trois mois. La moyenne et l'écart-type de la taille sont respectivement $\mu = 59,7$ et $\sigma = 3,2$. On va échantillonner p = 30 échantillons d'effectifs croissants (n = 8, 15, 20 puis 80) à partir de cette population (les résultats du premier échantillon sont ceux du Tableau 5.1). On a ainsi 4 distributions d'échantillonnage de la moyenne (Figure 5.3). On voit que (Figure 5.3 et Tableau 5.2) quand n augmente, la moyenne de la distribution d'échantillonnage se rapproche de μ avec de moins en moins de variabilité (σ_m tend vers 0).

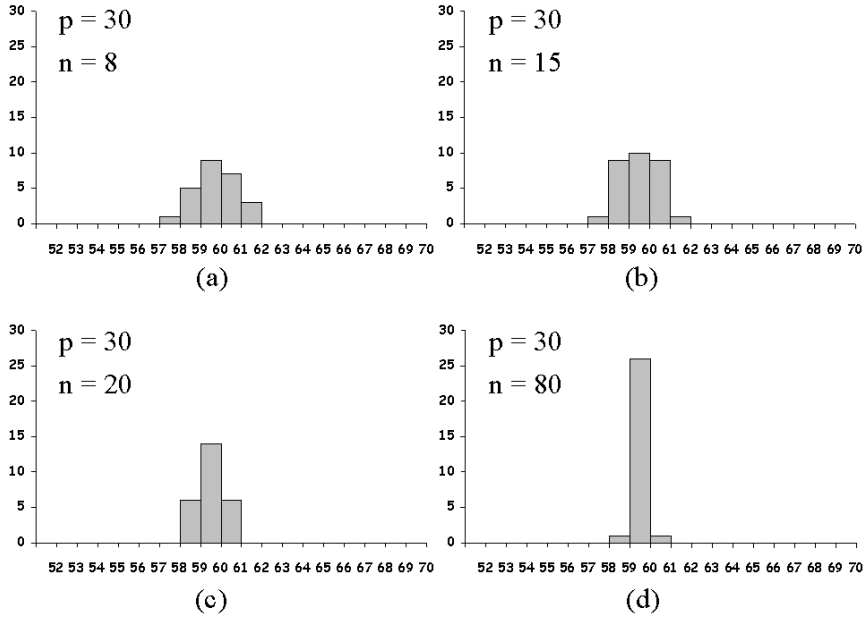

Figure 5.3 : Distributions d'échantillonnage de la moyenne pour p = 30 échantillons d'effectif (a) n = 8, (b) n = 15, (c) n = 20 et (d) n = 80.

							Population
n	8	15	20	80	…	↗	N = 200
m	59,8	59,6	59,6	59,7	…	59,7	μ = 59,7
σ_m	1,1	0,8	0,7	0,4	…	→ 0	σ = 3,2

Tableau 5.2 : Evolution de *m* et de σ_m en fonction de n.

Estimation de la variance d'une population

Soit x_1, x_2, …, x_n un échantillon tiré au hasard, d'effectif n et de moyenne $\bar{x} = \sum_{i=1}^{n} x_i \Big/ n$. L'**estimation de la variance de la population** est donnée par :

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

et s_x est une bonne estimation de l'écart-type de la population.

On peut démontrer que S_x^2 est un estimateur sans biais convergent de σ^2.

L'estimation de la variance de la v.a. $\bar{X}$, dont la loi de distribution est la loi de distribution d'échantillonnage de la moyenne, est donnée par :

$$s_m^2 = \frac{s_x^2}{n}$$

Une relation très utile permet de passer de la variance de l'échantillon x_1, x_2, …, x_n (Var(*X*)) à l'estimation de la variance de la population $\left(s_x^2\right)$:

$$s_x^2 = \frac{n}{n-1} \cdot \mathrm{Var}(X)$$

Exemple :

Supposons que l'on ait noté chez 11 individus normaux, pris au hasard, la valeur du rythme cardiaque. On a : 64, 80, 72, 88, 78, 88, 78, 88, 88, 72, 60.

On veut estimer la moyenne et la variance du rythme cardiaque chez les individus normaux (population cible). On a alors :

Estimation de la moyenne de la population (= moyenne de l'échantillon) :

$$\bar{x} = \frac{64+80+72+88+78+88+78+88+88+72+60}{11} = 77{,}82$$

Variance de l'échantillon :

$$\mathrm{Var}(X) = \frac{(64-77{,}82)^2 + \ldots + (60-77{,}82)^2}{11} = 90{,}51$$

Estimation de la variance de la population :

$$s_x^2 = \frac{(64-77{,}82)^2 + \ldots + (60-77{,}82)^2}{11-1} = 99{,}56$$

ou

$$s_x^2 = \frac{11}{11-1} \cdot 90{,}51 = 99{,}56$$

Estimation de l'écart-type de la population :

$$s_x = \sqrt{99{,}56} = 9{,}98$$

Estimation d'une proportion et de la variance d'une proportion (échantillon au hasard)

Exemple :

On applique un traitement à un groupe de 100 malades pris au hasard parmi tous les individus présentant la même affection. On observe 40 guérisons sur ce groupe. Le pourcentage de guérison observé pour cet échantillon de malades est de 40 %.

Estimation d'une proportion

Soit *k* le nombre de fois où un caractère donné est présent dans un échantillon tiré au hasard d'effectif *n* et soit *p* la proportion inconnue du caractère étudié dans la population.

La fréquence du caractère étudié dans l'échantillon vaut $f = k/n$, avec f qui est un observation de la v.a. F. On montre que :

$$\mathrm{E}(\mathrm{F}) = p$$

La fréquence d'un caractère étudié dans un échantillon tiré au hasard est une bonne estimation de la fréquence de ce caractère dans la population (F est sans biais).

Estimation de la variance d'une proportion

Par ailleurs :

$$\mathrm{Var}(\mathrm{F}) = \frac{p \cdot (1-p)}{n}$$

F est un estimateur convergent de *p*.

On estime la variance $p \cdot (1-p)/n$ par $f \cdot (1-f)/n$.

Estimation par intervalle

Définition

Nous avons vu, dans la partie « Estimation ponctuelle », que chacune des moyennes de *p* échantillons tirés au hasard $(\bar{x}_i,\ i = 1,..., p)$ est une estimation de la moyenne μ de la population et que chaque moyenne est différente d'un échantillon à l'autre. Ceci se généralise à la situation de l'estimation d'une proportion d'une population.

Notons θ un paramètre inconnu (une moyenne ou une proportion) d'une population. Si l'on souhaite que l'inférence réalisée à partir de $\hat{\theta}$ (estimation de θ obtenue sur un échantillon) présente un degré de confiance acceptable il faut construire un intervalle d'estimation (appelé **intervalle de confiance**), c'est-à-dire un intervalle, déterminé à partir des données d'un échantillon, dans lequel on peut parier, avec un risque de se tromper qui soit acceptable, que se situe réellement θ dans la population cible.

Ce risque, noté α, est généralement pris à 5 % et correspond aux erreurs d'échantillonnages jugées acceptables.

L'intervalle de confiance de θ est de la forme :

$$\hat{\theta} - \text{erreur d'échantillonnage} \ ; \ \hat{\theta} + \text{erreur d'échantillonnage}$$

Interprétation d'un intervalle de confiance :

On accepte qu'il y ait α.100 chances sur cent de se tromper en disant que θ appartient à l'intervalle.

On accepte qu'il y ait (1 - α).100 chances sur cent de ne pas se tromper en disant que θ appartient à l'intervalle.

Propriétés d'un intervalle de confiance :

- Toutes choses égales par ailleurs il est d'autant plus large que α est petit ;
- Toutes choses égales par ailleurs il est d'autant plus étroit que *n* est grand.

Exemple :

On dispose de 100 échantillons tirés au hasard. Pour chacun, on calcule la moyenne et l'intervalle de confiance de la moyenne. En prenant un risque d'erreur de 5 %, 95 intervalles de confiance contiendront la vraie valeur moyenne de la population et 5 ne la contiendront pas (Figure 5.4).

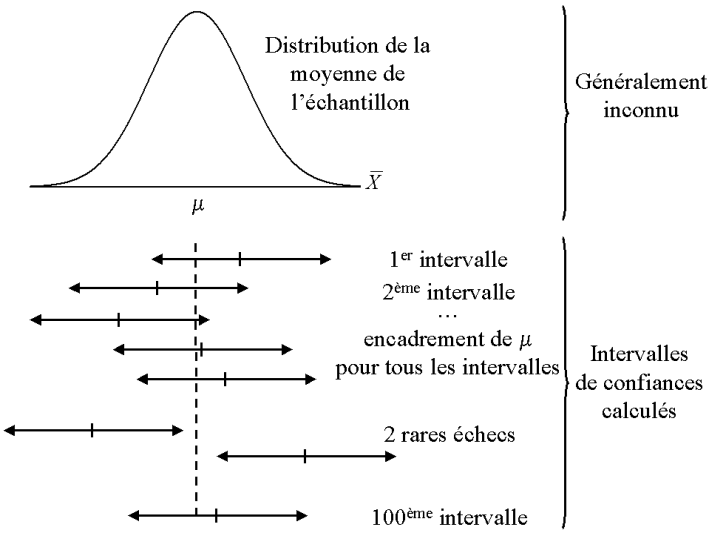

Figure 5.4 : Construction de 100 estimations d'intervalle. La vraie valeur μ est correctement encadrée dans 95 % des situations[3].

Intervalle de confiance d'une moyenne (échantillon au hasard)

Nous nous placerons dans le **cas où σ est inconnu**, ce qui est généralement le cas. On estimera alors σ par s_x.

Pour calculer l'intervalle de confiance d'une moyenne il est nécessaire de connaître la loi de distribution des moyennes des échantillons ou plus exactement loi de distribution de la quantité :

$$\frac{\bar{X}-\mu}{\frac{s_x}{\sqrt{n}}} = \frac{\bar{X}-\mu}{s_m}$$

Nous pouvons alors distinguer quatre situations selon que la variable d'origine suit ou non une loi Normale et selon la taille de l'échantillon étudié (grand, n ≥ 30, ou petit, n < 30). Les résultats mathématiques sur les lois de distributions dans ces quatre situations sont donnés dans le Tableau 5.3.

Les bornes de l'intervalle de confiance seront calculées pour chacune des situations en fonction de la loi de distribution des moyennes et du risque α choisi.

Ainsi, la forme générale de l'intervalle de confiance de la moyenne, calculé à partir de $\bar{x}$, estimation de la moyenne, de s_m, estimation de l'écart-type de la distribution des moyenne, et de la valeur L_α lue dans la table de la loi de distribution appropriée, est :

$$\left[\bar{x} - L_\alpha \cdot s_m \;;\; \bar{x} + L_\alpha \cdot s_m\right]$$

[3] Adapté de Wannacott & Wannacott. L'estimation par intervalle (chap. 8) in : Statistique : Economie, Gestion, Sciences, Médecine. Ed Economica, 1991.

Cas des grands échantillons (n ≥ 30)

Loi de distribution de la moyenne des échantillons : loi Normale (cf. Tableau 5.3).

Les bornes de l'intervalle de confiance, pour un risque α choisi, sont calculées par la formule :

$$\bar{x} \pm (N_\alpha \cdot s_m)$$

où N_α est la valeur lue dans la table de la loi Normale au risque α choisi.

Cas des petits échantillons (n < 30)

- Si la loi de distribution de la variable dans la population est Normale alors, la loi de distribution de la moyenne des échantillons est la loi de Student à $\nu = (n - 1)$ degrés de liberté (cf. Tableau 5.3).

  Les bornes de l'intervalle de confiance, pour un risque α choisi, sont calculées par la formule :

  $$\bar{x} \pm (T_{\alpha,\nu} \cdot s_m)$$

  où $T_{\alpha,\nu}$ est la valeur lue dans la table de la loi de Student au risque α choisi et avec $\nu = (n - 1)$ degrés de liberté.

- Si la loi de distribution de la variable dans la population n'est pas Normale alors on ne peut pas calculer l'intervalle de confiance des paramètres de la population cible.

	Hypothèse sur la loi de distribution de la variable dans la population	
	Suit une loi Normale N(μ, σ)	**Ne suit pas une loi Normale**
Petits échantillons (n < 30)	$\frac{\bar{X} - \mu}{\frac{s_x}{\sqrt{n}}}$ est une observation d'une loi de Student	On ne peut rien dire
Grands échantillons (n ≥ 30)	$\frac{\bar{X} - \mu}{\frac{s_x}{\sqrt{n}}}$ est une observation d'une loi de Student qui est proche de la loi N(0, 1)	$\frac{\bar{X} - \mu}{\frac{s_x}{\sqrt{n}}}$ est approximativement une observation d'une loi N(0, 1)

Tableau 5.3 : Détermination des lois de la v.a. étudiée en fonction des hypothèses de normalité et de la taille des échantillons. La case « On ne peut rien dire » correspond à une situation où

l'on ne peut rien dire sur la loi de distribution.

Exemple :

Sur un échantillon d'effectif n = 10 représentatif d'une population la moyenne $\bar{x} = 14$ *et l'estimation de l'écart-type de la population* $s_x = 2$.

Trouver l'intervalle de confiance, au risque de 5 % de la moyenne μ *de la population cible. On suppose que la variable suit une loi Normale.*

Il s'agit d'un petit échantillon (n < 30). La condition de Normalité étant remplie la loi de distribution de la moyenne des échantillons est la loi de Student à ν *degrés de liberté.*

n = 10, $\bar{x} = 14$, $s_x = 2$, *donc* $s_m = 2/\sqrt{10} = 0{,}63$.

Au risque $\alpha = 5\ \%$ *et pour* $\nu = 10 - 1 = 9$ *degrés de liberté,* $T_{0{,}05,\ 9} = 2{,}26$ *(valeur lue dans la table de Student).*

L'intervalle de confiance à 95 % de μ *est* $[14 - 2{,}26 \cdot 0{,}63 ; 14 + 2{,}26 \cdot 0{,}63] = [12{,}57 ; 15{,}43]$. *Autrement dit, on a 5 % de risque de se tromper en affirmant que l'intervalle [12,57 ; 15,43] recouvre* μ.

Intervalle de confiance d'une proportion (échantillon au hasard)

Si ***n* est suffisamment grand** et si ***f = k / n* n'est pas voisin de 1 ou de 0**, *k* étant le nombre de fois où le caractère donné est présent dans l'échantillon, on peut considérer que la distribution des fréquences F a une distribution Normale de moyenne *p* et d'écart-type s_f, avec :

$$s_f = \sqrt{\frac{f \cdot (1 - f)}{n}}$$

Ainsi, l'intervalle de confiance du pourcentage, pour un risque α choisi, calculé à partir de ***f*,** l'estimation de la fréquence, de s_f, l'écart-type de la distribution des fréquences, et de la valeur N_α lue dans la table de la loi Normale, est :

$$\left[f - N_\alpha \cdot s_f \ ; \ f + N_\alpha \cdot s_f \right]$$

Exemple :

Supposons que nous souhaitions estimer la fréquence d'une maladie que nous savons être comprise entre 20 % et 30 %.

Nous observons 12 malades sur un échantillon tiré au hasard de taille n = 48. L'intervalle de confiance à 95 % est (la table de la loi Normale nous donne $N_{0{,}05} = 1{,}96$*) :*

$$\left[\frac{12}{48} \pm \left(1{,}96 \cdot \sqrt{\frac{\frac{12}{48} \cdot \frac{36}{48}}{48}} \right) \right] = \left[0{,}25 \pm \left(1{,}96 \cdot \sqrt{\frac{\frac{1}{4} \cdot \frac{3}{4}}{48}} \right) \right] = \left[0{,}25 \pm \left(1{,}96 \cdot \sqrt{\frac{3}{16 \cdot 16 \cdot 3}} \right) \right]$$

soit,

$$0{,}25 \pm \frac{2}{16} = 0{,}25 \pm 0{,}125 = [12{,}5\% \ ; \ 37{,}5\%]$$

ce qui n'apporte aucune information (l'échantillon est de trop petite taille).

Supposons que nous souhaitions obtenir un intervalle de confiance à 95 % de demi largeur (précision) 2 %.

On peut penser que la fréquence observée sera comprise entre 0,2 et 0,3 ; on peut par exemple supposer que cette fréquence sera de 0,2.

L'intervalle de confiance sera donc : $\left[0{,}2 \pm \left(1{,}96 \cdot \sqrt{\frac{0{,}2 \cdot 0{,}8}{n}}\right)\right]$

Or on veut que $\left[\left(1{,}96 \cdot \sqrt{\frac{0{,}2 \cdot 0{,}8}{n}}\right) \leq \frac{2}{100}\right]$

Soit, en arrondissant 1,96 à 2 :

$$2 \cdot \sqrt{\frac{0{,}2 \cdot 0{,}8}{n}} \leq \frac{2}{100} \Leftrightarrow n \geq \frac{4 \cdot 0{,}2 \cdot 0{,}8}{4} \cdot 100^2 = 1600$$

Il faut donc pour espérer obtenir un intervalle informatif (étroit) prendre un échantillon tiré au hasard de taille supérieure ou égale à 1600.

Supposons que nous observions 400 malades sur 1600 ; l'intervalle de confiance à 95 % est :

$$\left[0{,}25 \pm \left(1{,}96 \cdot \sqrt{\frac{\frac{1}{4} \cdot \frac{3}{4}}{1600}}\right)\right] = [0{,}25 \pm 0{,}021] = [22{,}9\% \ ; \ 27{,}1\%]$$

Remarque :

Exprimer les résultats d'un échantillon en indiquant uniquement sa moyenne, ou la fréquence du caractère étudié, est sans valeur. Elle ne suffit pas à caractériser l'ensemble des mesures effectuées. Il manque une information sur la dispersion des mesures observées et le nombre de mesures sur lequel ces calculs ont été effectués. Il faut donc exprimer les résultats d'une série de mesures en indiquant son paramètre de position et un paramètre de dispersion (cf. Tableau 5.4).

Situation d'estimation d'une	
Moyenne	**Proportion**
La moyenne $\bar{x}$	La proportion f

La variance estimée (ou l'écart-type) : s_x^2	La variance estimée (ou l'écart-type) : s_f^2
nombre de mesures effectuées : n	nombre de mesures effectuées : n

Tableau 5.4 : Expression des résultats relatifs à un échantillon.

Ce qu'il faut savoir absolument

Les caractéristiques de la **population** cible sont généralement inconnues. Un **échantillonnage** par **tirage au sort aléatoire** permet de constituer un **échantillon représentatif** de la population cible. On obtient à partir des données provenant de l'échantillon une **estimation** d'une caractéristique inconnue de la population.

Qualité d'un estimateur : un bon estimateur est sans biais avec une variance qui tend vers 0 quand l'effectif de l'échantillon observé tend vers l'infini (il est alors **convergent**).

Estimation de la moyenne et de la variance d'une population :

Soit x_1, x_2, …, x_n un échantillon tiré au hasard d'effectif n et de moyenne

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

alors l'observation $\bar{x}$ de $\bar{X}$ est une bonne estimation de la moyenne μ de la population cible et

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

est une bonne estimation de la variance de la population et sa racine carrée, s_x, est une bonne estimation de l'écart-type de la population.

La formule suivante permet d'avoir une estimation de la variance de la population à partir de la variance de l'échantillon :

$$s_x^2 = \frac{n}{n-1} \cdot \mathrm{Var}(X)$$

Estimation d'une proportion et de la variance d'une proportion :

La fréquence du caractère étudié dans l'échantillon tiré au hasard, $f = k/n$, est une bonne estimation de la fréquence de ce caractère dans la population et

$f \cdot (1-f)/n$ est une bonne estimation de la variance de la fréquence de ce caractère dans un échantillon de taille n.

Intervalle de confiance d'une moyenne

	Hypothèse sur la loi de distribution de la variable dans la population	
	Suit une loi Normale	**Ne suit pas une loi Normale**
Petits échantillons (n < 30)	$\left[\bar{x} - T_{\alpha,\nu} \cdot s_m \;;\; \bar{x} + T_{\alpha,\nu} \cdot s_m\right]$	On ne peut rien dire
Grands échantillons (n ≥ 30)	$\left[\bar{x} - N_\alpha \cdot s_m \;;\; \bar{x} + N_\alpha \cdot s_m\right]$	

avec $s_m = \frac{s_x}{\sqrt{n}}$

Intervalle de confiance d'une proportion

$$\left[f - N_\alpha \cdot s_f \;;\; f + N_\alpha \cdot s_f\right]$$

avec $s_f = \sqrt{\frac{f \cdot (1-f)}{n}}$

Interprétation d'un intervalle de confiance :

Soit θ le paramètre inconnu (une moyenne ou une proportion) d'une population que l'on cherche à estimer.

On dira qu'il y a α.100 chances sur cent de se tromper en disant que θ appartient à l'intervalle.

On dira qu'il y a (1 - α).100 chances sur cent de ne pas se tromper en disant que θ appartient à l'intervalle.

Propriétés d'un intervalle de confiance :

- Il est centré sur l'estimation du paramètre ;
- Toutes choses égales par ailleurs il est d'autant plus large que α est petit ;
- Toutes choses égales par ailleurs il est d'autant plus étroit que n est grand.

Chapitre 6 Principes généraux des tests statistiques

Position du problème (exemple)

Dans une industrie pharmaceutique, une machine fabrique des gélules. Elle est réglée de telle sorte que chaque gélule contienne une quantité μ de produit actif. En réalité la quantité de produit actif dans une gélule est aléatoire et nous admettrons que cette quantité suit une loi $N(\mu,\sigma)$, σ étant connu. On voudrait savoir si la machine est réglée correctement et délivre bien la quantité 'a' *de produit actif. Pour cela, on prélève un échantillon au hasard de taille n et pour chacune des gélules on mesure (sans erreur) la quantité de produit actif. Il s'agira, au vu de la quantité moyenne de produit actif observée, $\bar{x}$, de dire si le réglage est bon ou mauvais. Si le réglage est bon, la quantité moyenne de produit actif est toujours μ =* a. *Si le réglage est mauvais, la quantité moyenne de produit actif est μ, différente de* 'a'. *On doit donc réaliser un test de comparaison pour comparer une moyenne observée à une constante.*

Cependant, pour chacune de ces situations, la quantité moyenne de produit actif observée sur un échantillon tiré au hasard pourra prendre une valeur différente de la valeur théorique (Tableau 6.1). En effet, toute mesure effectuée sur un échantillon est soumise aux ***fluctuations d'échantillonnage****, dues au hasard du tirage.*

	Réglage bon	**Réglage mauvais**
Quantité moyenne de produit actif théorique	μ = a	$\mu \neq$ a
Quantité moyenne de produit actif observée sur échantillon	$\bar{x} \approx$ a	$\bar{x} \neq$ a
Valeurs possibles de $\bar{x}$ observées sur échantillon	0 → infini	0 → infini

Tableau 6.1 : Effets des fluctuations d'échantillonnage.

On ne peut donc dire avec certitude si le réglage est bon ou mauvais. On répondra alors à cette question en acceptant un certain risque d'erreur.

Intuitivement, on dira que le réglage est mauvais si l'écart entre $\bar{x}$ *et* 'a' *est « grand ». Un test de comparaison permet d'associer au qualificatif subjectif « grand » un risque d'erreur connu et accepté.*

Méthode « classique » d'un test statistique

Au cours d'une expérience, on a prélevé un échantillon au hasard de 100 gélules. On a observé une quantité moyenne de produit actif, mesurée sans erreur, $\bar{x}$. Rappelons que la machine est supposée être réglée de telle sorte que la quantité de produit actif dans une gélule vaut 'a'. Pour savoir si le réglage est bon ou mauvais on réalise un test de comparaison. Pour cela, plusieurs étapes sont nécessaires :

1. Définir l'hypothèse H_0 que l'on cherche à tester, classiquement appelée **hypothèse nulle**. Dans ce cas, on veut tester si la quantité moyenne de produit actif théorique, μ, est égale à la quantité ordonnée 'a', ou plus simplement, si leur différence est nulle : H_0 : μ - a = 0, c'est-à-dire que le réglage est bon. Par la suite, nous noterons D la variable aléatoire correspondant à cette différence et $d = \bar{x} - a$ sa valeur observée.
2. Fixer le **risque d'erreur** global acceptable du test dans l'hypothèse où H_0 est vraie. Ce risque, dorénavant noté α, est usuellement fixé à 5 %.
3. Supposons que l'hypothèse nulle soit vraie ; alors la distribution de la différence théorique D est connue (Figure 6.1). Le mauvais réglage de la machine pouvant modifier la quantité moyenne de produit actif en l'augmentant ou en la diminuant, le risque d'erreur (α = 5 %) se décompose en 2 (zones grisées Figure 6.1). Ce risque d'erreur ainsi que les propriétés de la loi de la statistique D lorsque H_0 est vraie déterminent une valeur seuil $|V_s|$ telle que les valeurs supérieures à $|V_s|$ seront jugées trop éloignées de 0 (d est « grand ») avec un risque d'erreur global consenti de α.
4. Calculer la statistique du test de comparaison d'une moyenne observée à une constante. Si cette valeur appartient à une des régions de rejet, alors on considère que cette valeur est suffisamment en désaccord avec H_0 pour rejeter cette dernière ; sinon, H_0 n'est pas rejetée.

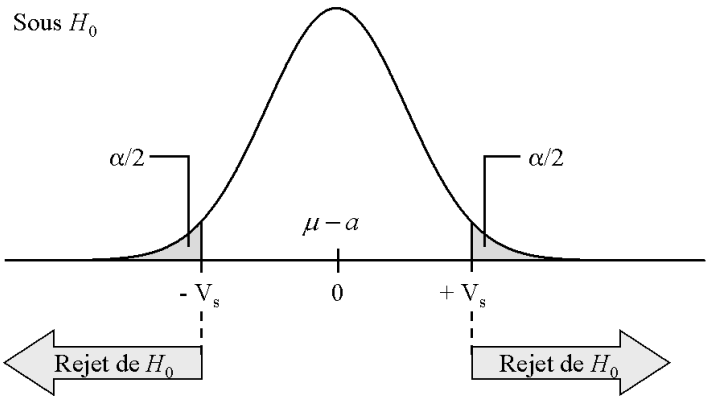

Figure 6.1 : Représentation d'un test statistique.

Dans la méthode classique, la conclusion au test statistique repose sur la comparaison entre la valeur du résultat de la statistique du test et la valeur seuil :

> Rejet de H_0 si : |résultat de la statistique du test| ≥ |valeur seuil|
>
> Conservation de H_0 si : |résultat de la statistique du test| < |valeur seuil|

Notion de risque

Nous avons vu que la réalisation d'un test implique de définir une hypothèse nulle H_0 que l'on veut tester. La Figure 6.2 (a) représente ce que l'on observe réellement, si H_0 est vraie. On appelle **seuil de signification** la valeur V_s correspondant au risque de rejeter H_0 alors que celle-ci est vraie. Ce type d'erreur s'appelle également **erreur de 1ère espèce** et correspond au **risque d'erreur α ou risque de 1ère espèce**,.

Le rejet de l'hypothèse H_0 se fait au bénéfice d'une autre hypothèse, dite **hypothèse alternative (H_A)**. Comme sous H_0, D suit sous H_A une loi de distribution dont les caractéristiques peuvent être étudiées. Une erreur peut être commise si le résultat de la statistique du test de comparaison tombe dans la zone d'acceptation de H_0, alors que H_A est vraie. On appelle **erreur de 2ème espèce** l'erreur consistant à accepter H_0 alors que celle-ci est fausse. Sa probabilité β est le **risque de 2ème espèce**, représentée par la zone rayée sur la Figure 6.2 (b).

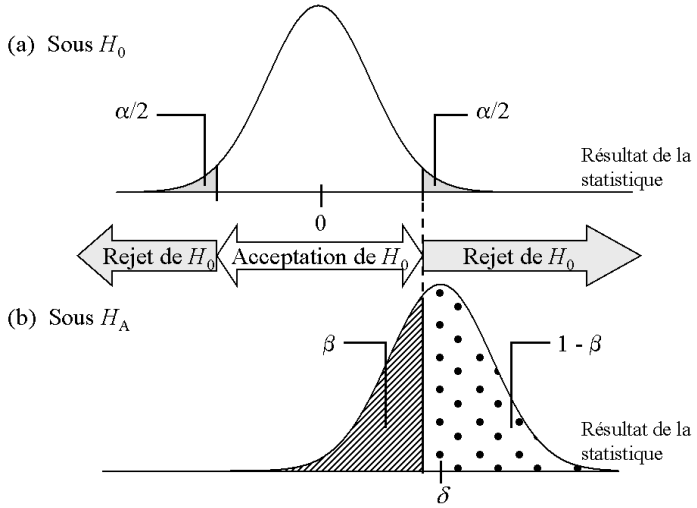

Figure 6.2 : Types d'erreurs possibles dans un test. (a) α = probabilité de rejeter H_0 alors qu'elle est vraie (b) β = probabilité d'accepter H_0 alors qu'elle est fausse. δ est le résultat de la statistique.

Une décision de rejet ou d'acceptation d'une hypothèse est toujours prise avec incertitude (la réalité est inconnue). Le Tableau 6.2 représente ces différentes situations décisionnelles.

		Réalité (inconnue)	
		H_0 vraie	H_A vraie
Décision retenue au vue du résultat de la statistique	H_0 vraie	Pas d'erreur	Risque β
	H_A vraie	Risque α	Pas d'erreur

Tableau 6.2 : Risques de première espèce (α) et de deuxième espèce (β).

A côté de ces situations décisionnelles d'erreurs, une situation décisionnelle correcte est particulièrement intéressante : la **puissance d'un test** (**1 - β**) est la capacité de ce test à montrer une différence si elle existe (représentée par la zone à pois sur la Figure 6.2 (b)).

D'une manière générale, pour la même hypothèse nulle d'égalité des moyennes, il est possible de considérer 2 types d'hypothèse alternative :

- **Test bilatéral** (Figure 6.3) **:** soit une situation consistant à comparer 2 moyennes μ_0 et μ_A et où le sens de la différence importe peu. H_0 correspond à $\mu_0 = \mu_A$ et H_A admet alors aussi bien $\mu_0 > \mu_A$ que $\mu_0 < \mu_A$. On veut savoir si les deux moyennes sont statistiquement significativement différentes ou non, sans s'occuper de savoir laquelle est supérieure à l'autre ($H_A : \mu_0 \neq \mu_A$).

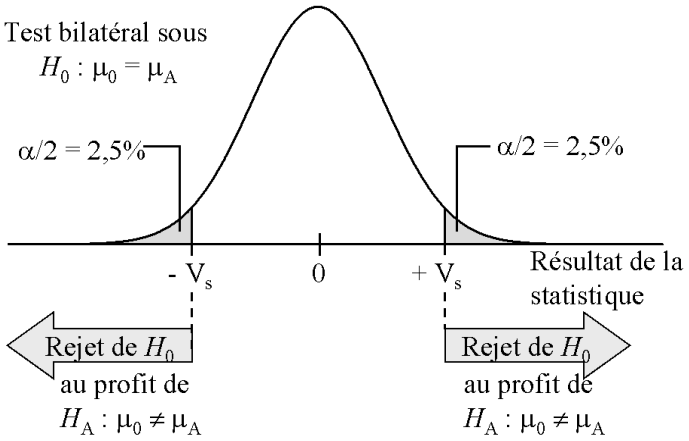

Figure 6.3 : Test bilatéral.

- **Test unilatéral** (Figure 6.4) **:** une autre situation consiste à introduire une notion d'ordre dans la définition de H_A : la moyenne μ_0 est inférieure à la moyenne μ_A (ou l'inverse, $H_A : \mu_0 > \mu_A$). Si $H_A : \mu_A > \mu_0$, seules les valeurs du résultat de la statistique supérieures à $+V_s$ sont en faveur de H_A et la région de rejet n'a qu'un seul côté. D'une manière similaire à la situation bilatérale, V_s est déterminé par : Proba($D \geq +V_s$) = α.

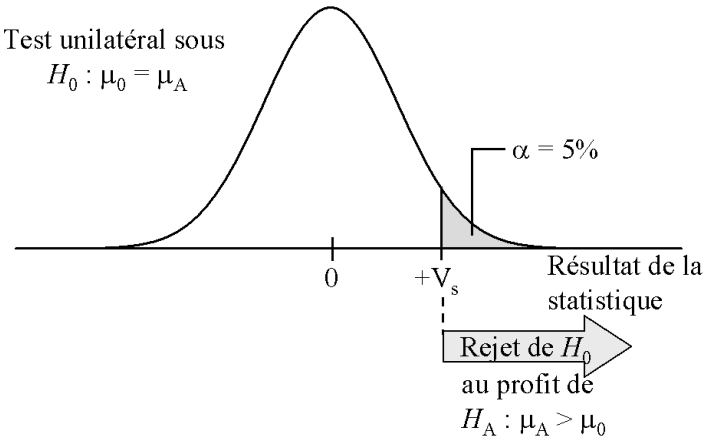

Figure 6.4 : Test unilatéral.

Degré de signification d'un test statistique

Nous avons vu comment conclure à un test en comparant le résultat de la statistique avec la valeur seuil, déterminée par α et la loi de la statistique lorsque l'hypothèse nulle est vraie. Le résultat de cette procédure (rejet de H_0 ou acceptation de H_0) s'exprime avec un risque d'erreur α fixé *a priori* et arbitrairement. Il existe une autre approche qui permet de quantifier la crédibilité de H_0 au vue des données observées. Elle repose sur le calcul du **degré de signification,** noté **p,** qui est la probabilité d'observer une différence au moins aussi importante que celle observée, sous l'hypothèse nulle :

p = Proba(valeur de la statistique ≥ valeur calculée de la statistique si H_0 est vraie)

Comme nous le verrons, il suffit d'utiliser les tables habituelles des lois de distributions pour obtenir une valeur approchée de p.

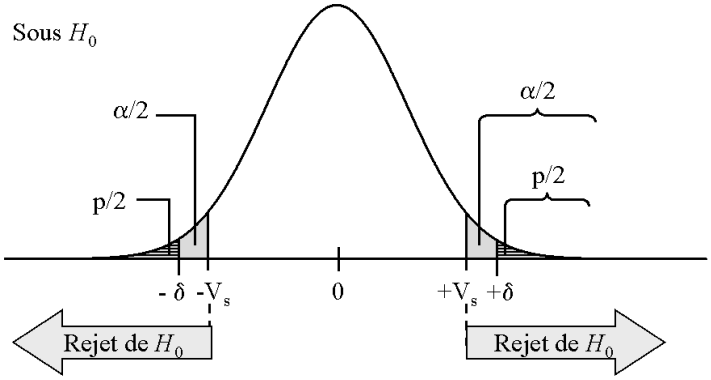

Figure 6.5 : Représentation du risque d'erreur α et du degré de signification (p) sous H_0.

La Figure 6.5 représente, sous H_0, le risque d'erreur α et le degré de signification pour une valeur calculée δ de la statistique.

Dans la méthode basée sur le degré de signification, la conclusion au test statistique repose sur la comparaison entre la valeur du degré de signification et la valeur de α :

> Rejet de H_0 si : $p \leq \alpha$
>
> Conservation de H_0 si : $p > \alpha$

Habituellement, on conclut selon la méthode classique, avec en général un risque d'erreur $\alpha = 5\%$, et on donne le degré de signification p.

Remarques 1 :

- $p \leq \alpha \Leftrightarrow$ valeur calculée de la statistique ≥ valeur seuil ;
- valeur calculée de la statistique ↘ ⇒ ↗ p

Remarques 2 : Interprétations erronées de p :

- Il est faux de dire que « p est le risque ou la probabilité de rejeter à tort l'hypothèse nulle ».

En effet p est la traduction en terme de probabilité de la valeur observée de la

statistique, et est donc une valeur observée ; p traduit simplement en terme de probabilité l'éloignement entre la valeur observée de la statistique et la valeur attendue sous l'hypothèse nulle.

- Interpréter p en terme de « force de la différence » **est abusif.**

Une conclusion du type : « les durées moyennes de séjour diffèrent fortement (p = 0,0001) » est abusive. Pour apprécier l'écart entre les durées de séjour il convient de donner un intervalle de confiance de la différence des durées moyennes. En effet p peut être petit parce que l'écart entre la réalité et l'hypothèse nulle est grand, parce que la puissance est élevée, ou les deux, il se peut aussi que p soit petit « par hasard » et même par erreur de 1ère espèce.

Variations de β

Variation de β en fonction de α

Supposons que le même test soit réalisé d'une part au risque α_1 et d'autre part au risque α_2, avec $\alpha_2 < \alpha_1$. Comme le montre de manière intuitive la Figure 6.6, une diminution du risque α augmente la valeur seuil ($V_{s2} > V_{s1}$) entraînant de ce fait une augmentation du risque β ($\beta_2 > \beta_1$).

Toute chose égale par ailleurs, **β et α varient en sens inverse**.

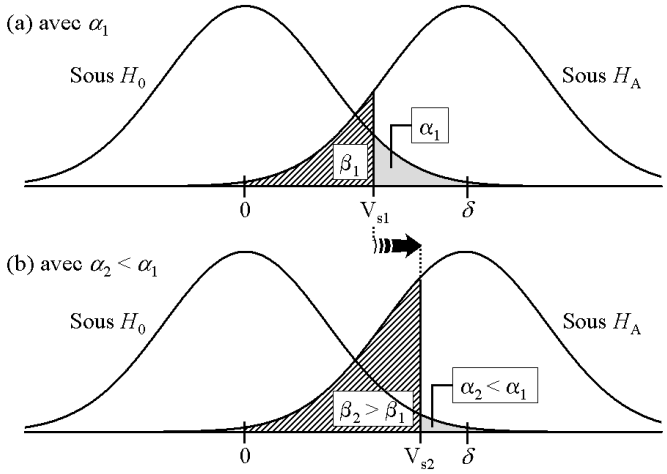

Figure 6.6 : Variation de β en fonction de α (toute chose égale par ailleurs). (a) Situation pour un risque α_1 fixé. (b) Situation pour un risque $\alpha_2 < \alpha_1$ (V_s correspond à la valeur seuil et δ correspond au résultat de la statistique).

Variation de β en fonction de la taille de l'échantillon

La précision d'une estimation augmente avec la taille de l'échantillon (cf. le chapitre « Estimation »), ce qui se traduit graphiquement par un resserrement de la courbe de distribution autour de la valeur estimée. Ainsi, sous H_0, la courbe de la distribution de la différence théorique des moyennes d'un échantillon de taille $n_2 > n_1$ se resserre autour de 0 et, pour conserver un risque $\alpha = 5$ %, la valeur seuil du test diminue

(Figure 6.7). La courbe de la distribution sous H_A se resserrant également du fait de l'augmentation de n, il s'en suit une diminution de β.

Toute chose égale par ailleurs, **β et n varient en sens inverse**.

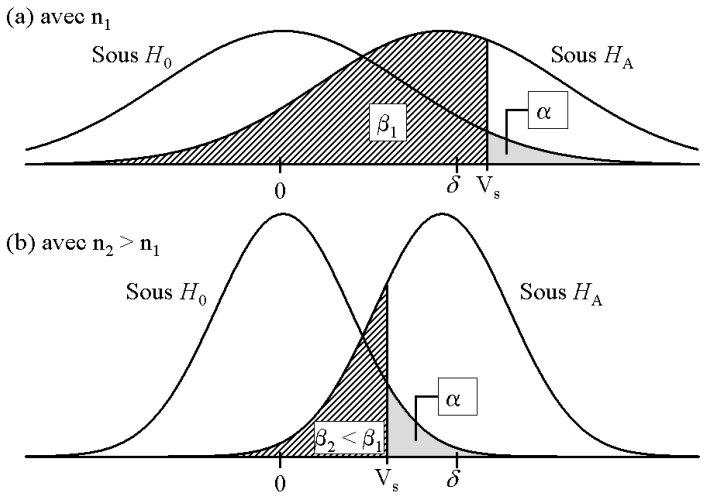

Figure 6.7 : Variation de β en fonction de la taille de l'échantillon (toute chose égale par ailleurs). (a) Situation pour un effectif n_1 fixé. (b) Situation pour un effectif $n_2 > n_1$ (V_s correspond à la valeur seuil et δ correspond au résultat de la statistique).

Variation de β en fonction de l'écart H_0 - H_A

Supposons que 2 tests aient été réalisés et que l'écart entre H_0 et H_A pour le 2ème test soit supérieur à l'écart entre H_0 et H_A pour le 1er test. Alors, toute chose égale par ailleurs, le risque β du 2ème test sera inférieur au risque β du 1er test.

Raisonner à partir de l'écart entre le résultat de la statistique du test, noté δ, et 0 (test de comparaisons d'une moyenne théorique à une moyenne observée, comme nous l'avons vu) est équivalent au raisonnement à partir de l'écart H_0 - H_A. Comme le montre la Figure 6.8, l'accroissement de la valeur δ éloigne la distribution sous H_A de la distribution sous H_0 et, toute chose égale par ailleurs, conduit à une diminution de β.

Toute chose égale par ailleurs, **β et l'écart H_0 - H_A varient en sens inverse**.

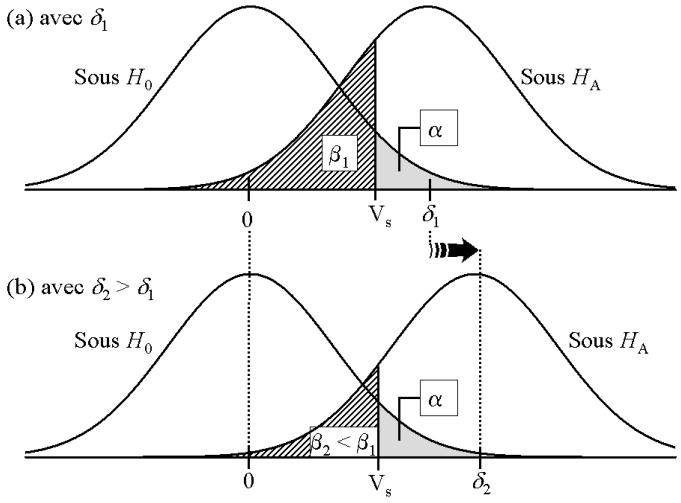

Figure 6.8 : Variation de β en fonction de l'écart H_0 - H_A (toute chose égale par ailleurs). (a) Situation pour un résultat de la statistique δ_1 fixé. (b) Situation pour un résultat $\delta_2 > \delta_1$ (V_s correspond à la valeur seuil).

Récapitulatif

Le Tableau 6.3 récapitule l'interdépendance entre β, et donc entre la puissance d'un test (1 - β), et certains paramètres statistiques. Ces propriétés sont vraies pour tous les tests statistiques.

Puissance augmente (β diminue)	Si α augmente
	Si n augmente
	Si Δ (écart H_0 - H_A) augmente

Tableau 6.3 : Variation de la puissance d'un test statistique.

Choix d'un test statistique

Pour choisir le « bon » test statistique il est nécessaire de porter une réflexion sur les données du problème à analyser. Il existe différents points, communs à tous les tests statistiques présentés par la suite dans ce polycopié, qu'il convient de relever :

1. **Types de variables** mises en relation :

S'agit-il d'étudier le lien entre deux variables de type quantitatif ?

Est-ce que l'on a à comparer entre elles deux variables de type qualitatif ?

Compare t-on une variable de type quantitatif à une variable de type qualitatif ?

Le type des variables déterminera le test statistique, ou un ensemble de tests statistiques, qui pourront être utilisables (sous certaines conditions qui leur sont propres).

2. **Taille de l'échantillon** :

L'échantillon est-il de taille < 30 ou ≥ 30 ? Cette limite permet alors dans la deuxième situation d'utiliser une statistique de test fondée sur une distribution Normale.

3. **Conditions d'applications** des tests choisis :

Les tests statistiques sont utilisables dans certaines conditions bien définies et souvent spécifiques (par exemple, distribution Normale dans la population). Ces conditions seront spécifiées lors de la présentation de chaque test.

4. **Séries non appariées ou appariées** :

Dans les séries non appariées les comparaisons portent sur des observations provenant d'individus différents pris au hasard indépendamment. Ces observations sont alors indépendantes entre elles.

Dans les séries appariées les comparaisons portent sur des observations qui ne sont pas indépendantes. C'est la cas, par exemple, des expériences du type « mesure avant - mesure après » qui intéressent les mêmes individus dans les deux échantillons.

Exemple :

On a mesuré la fibrinémie avant un traitement A et après ce même traitement. On dispose donc pour chaque individu du dosage de la fibrine à deux instants donnés, avant et après le traitement. On a donc à comparer une variable de type quantitatif (dosage de la fibrine) à une variable de type qualitatif (dont les modalités sont « avant » et « après ») et l'on pourrait alors penser traiter ce problème en faisant un test de comparaison des moyennes de l'échantillon de mesures « avant » et de l'échantillon de mesures « après ». Mais cette solution ne serait pas exacte. En effet, on suppose dans un tel test que les deux échantillons sont indépendants. Or, dans notre exemple, les résultats des fibrinémies des deux séries se correspondent deux à deux, c'est-à-dire que chacun des nombres de la première série de mesures doit être comparé au nombre correspondant de la deuxième série de mesures et non à tous les autres. Ici, on veut connaître l'effet d'un changement de situation sur une variable mesurée chez les mêmes individus.

Les étapes d'un test statistique

1. Choix des hypothèses à tester :
   - Choix d'une hypothèse nulle H_0 ;
   - Choix d'une hypothèse alternative H_A (acceptée si H_0 est rejetée).
2. Fixer une règle (choix du risque α, habituellement 5 %) pour décider l'acceptation ou le rejet de H_0 ;
3. Vérification des conditions d'application :
   - Choix des échantillons au hasard dans les populations ;
   - Taille des échantillons et autres conditions spécifiées dans le chapitre suivant.
4. Calcul de la statistique appropriée ;
5. Décision en comparant la valeur de la statistique calculée au seuil de signification correspondant au risque α choisi ;
6. Calcul du degré de signification du test.

Ce qu'il faut savoir absolument

La conclusion d'un test statistique est soit la conservation de l'**hypothèse nulle** soit l'acceptation de l'**hypothèse alternative** selon le résultat de la comparaison de la valeur calculée de la statistique avec la valeur seuil. Cette conclusion s'accompagne du **degré de signification** du test qui est la probabilité d'observer une différence au moins aussi importante que celle observée, sous l'hypothèse nulle.

Notion de risque

On appelle **seuil de signification** la valeur V_s correspondant au risque de rejeter l'hypothèse nulle alors que celle-ci est vraie. Ce type d'erreur s'appelle également **risque d'erreur α** ou **risque de 1ère espèce**. Il y a une autre façon de se tromper : on peut accepter l'hypothèse nulle alors que l'hypothèse alternative est vraie. Ce risque est appelé **risque de 2ème espèce** ou **risque d'erreur β**.

- Une erreur de première espèce est commise si on rejette H_0 alors que celle-ci est vraie (rejet de H_0 à tord).
- Une erreur de deuxième espèce est commise si on accepte H_0 alors que celle-ci est fausse (acceptation de H_0 à tord).

		Réalité (inconnue)	
		H_0 vraie	H_A vraie
Décision retenue au vue du résultat de la statistique	H_0 vraie	Pas d'erreur	Risque β
	H_A vraie	Risque α	Pas d'erreur

Figure 6.9 : Risques de première (α) et de deuxième espèce (β).

Puissance et tests statistiques

Puissance augmente (β diminue)	Si α augmente
	Si n augmente
	Si Δ (écart H_0 - H_A) augmente

Figure 6.10 : Variation de la puissance d'un test statistique.

Annexe : Tables utiles

Table de la loi Normale

α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	∞	2,58	2,33	2,17	2,05	**1,96**	1,88	1,81	1,75	1,69
0,10	1,64	1,60	1,55	1,51	1,48	1,44	1,40	1,37	1,34	1,31
0,20	1,28	1,25	1,23	1,20	1,17	1,15	1,13	1,10	1,08	1,06
0,30	1,04	1,01	0,99	0,97	0,95	0,93	0,91	0,90	0,88	0,86
0,40	0,84	0,82	0,81	0,79	0,77	0,75	0,74	0,72	0,71	0,69
0,50	0,67	0,66	0,64	0,63	0,61	0,60	0,58	0,57	0,55	0,54
0,60	0,52	0,51	0,50	0,48	0,47	0,45	0,44	0,43	0,41	0,40
0,70	0,38	0,37	0,36	0,34	0,33	0,32	0,30	0,29	0,28	0,27
0,80	0,25	0,24	0,23	0,21	0,20	0,19	0,18	0,16	0,15	0,14
0,90	0,13	0,11	0,10	0,09	0,07	0,06	0,05	0,04	0,02	0,01

Table de la loi de Student

Nombre de d.d.l.	$\alpha = 0,20$	$\alpha = 0,10$	$\alpha = 0,05$	$\alpha = 0,02$	$\alpha = 0,01$
1	3,078	6,314	12,706	31,821	63,66
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
≥ 30	1,28	1,64	**1,96**	2,33	2,58

Table de la loi du Chi2

nombre de d.d.l.	α = 0,30	α = 0,20	α = 0,10	α = 0,05	α = 0,01	α = 0,001
1	1,07	1,64	2,71	**3,84**	6,63	10,83
2	2,41	3,22	4,60	5,99	9,21	13,81
3	3,66	4,64	6,25	7,81	11,34	16,27
4	4,88	5,99	7,78	9,49	13,28	18,47
5	6,06	7,29	9,24	11,07	15,09	20,51
6	7,23	8,56	10,64	12,59	16,81	22,46
7	8,38	9,80	12,02	14,07	18,47	24,32
8	9,52	11,03	13,36	15,51	20,09	26,12
9	10,66	12,24	14,68	16,92	21,67	27,88
10	11,78	13,44	15,99	18,31	23,21	29,59

Table du coefficient de corrélation

nombre de d.d.l.	α = 0,10	α = 0,05	α = 0,01
1	0,988	0,997	0,999
2	0,900	0,950	0,990
3	0,805	0,878	0,959
4	0,729	0,811	0,917
5	0,669	0,754	0,874
6	0,621	0,707	0,834
7	0,582	0,666	0,798
8	0,549	0,632	0,765
9	0,521	0,602	0,735
10	0,497	0,576	0,708
11	0,476	0,553	0,683
12	0,457	0,532	0,661
13	0,441	0,514	0,641
14	0,426	0,497	0,623
15	0,412	0,482	0,605
16	0,400	0,468	0,590
17	0,389	0,455	0,575
18	0,378	0,444	0,561
19	0,369	0,433	0,549
20	0,360	0,423	0,537

Index