Histoire 1

Les statistiques, nées de la nécessité de comprendre les données et les phénomènes aléatoires, ont évolué depuis les recensements de l'Antiquité. Le mathématicien Karl Pearson, pionnier de la statistique moderne, a introduit des concepts clés comme le coefficient de corrélation. Ronald Fisher, un autre géant du domaine, a développé l'analyse de la variance, fondant ainsi la statistique inférentielle. Ensemble, ils ont façonné la discipline qui permet aujourd'hui d'interpréter les données avec rigueur.

1 Analyse statistique bivariée

1.1 Vocabulaire des statistiques

Définition 1

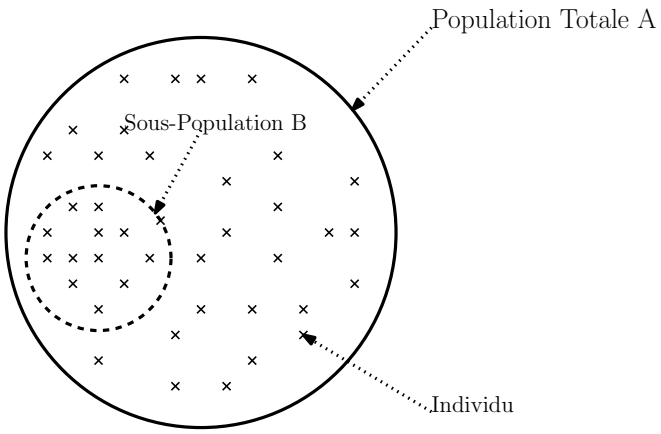
L'objet de la *statistique* est l'étude d'une *population*, formée d'*individus* sur lesquels on observe des *caractères*.
On étudie souvent une population humaine ; les individus sont des personnes, et les caractères observés peuvent être la taille, le poids, le groupe sanguin, le taux de cholestérol etc ... mais les notions de population et de caractère sont plus générales. Quelques exemples :

Population	Galaxie	Corps électoral	Ville	Pays
Caractère	Nombre d'étoiles	Abstentionniste	Taux d'imposition	PIB

Un caractère est dit :

- *qualitatif*, quand les valeurs ne peuvent être ni ordonnées ni ajoutées (groupe sanguin, couleur des yeux, vote pour un candidat)
- *quantitatif*, quand les valeurs sont numériques (mesures physiques, physiologiques, économiques).

Si on étudie deux caractères pour chaque individu de la population (par exemple l'âge et le sexe), on parle de *statistique bivariée*.



1.2 Tableau croisé d'effectifs

Capacité 1 *Calculer des effectifs marginaux à partir d'un tableau croisé d'effectifs*

Une usine fabrique des toiles de pergola. Ces toiles peuvent présenter deux types de défauts :

- Un défaut de couleur de la toile ;
- un défaut d'étanchéité de la toile.

Sur un lot de 200 toiles, on obtient les résultats présentés dans le tableau croisé d'effectifs ci-dessous :

	Présente un défaut d'étanchéité (E)	Ne présente pas de défaut d'étanchéité ($\overline{E}$)	Total
Présente un défaut de couleur (C)	8	8	…
Ne présente pas de défaut de couleur ($\overline{C}$)	4	180	…
Total	…	…	200

1. Dans cette étude statistique, déterminer les individus et les caractères étudiés.

2. Combien de toiles présentent un défaut d'étanchéité et de couleur ? On note $n(C \cap E)$ cet effectif.

3. Décrire en français l'ensemble $\overline{C} \cap E$ et déterminer son effectif $n(\overline{C} \cap E)$.

4. Calculer le nombre total de toiles présentant un défaut de couleurs. Compléter le tableau avec cet effectif marginal de C, noté $n(C)$.

5. Compléter le tableau avec les effectifs marginaux $n(\overline{C})$, $n(E)$ et $n(\overline{E})$.

Définition 2

Un **tableau croisé d'effectifs** (ou tableau à double entrée) est un tableau donnant les effectifs portant sur deux caractères d'une même population.
Les valeurs de l'un des caractères sont présentées en ligne et celles de l'autre caractère sont présentées en colonne.
Les sommes des colonnes et des lignes (nommées « Total ») sont aussi appelées les **marges** du tableau.

$C_1 \backslash C_2$	B	$\overline{B}$	Total
A	$n(A \cap B)$	$n(A \cap \overline{B})$	$n(A)$
$\overline{A}$	$n(\overline{A} \cap B)$	$n(A \cap \overline{B})$	$n(\overline{A})$
Total	$n(B)$	$n(\overline{B})$	N

Le tableau croisé d'effectifs ci-contre présente les effectifs du caractère C_1 de valeurs A et $\overline{A}$ (en ligne) et du caractère C_2 de valeurs B et $\overline{B}$ (en colonne).
$n(A \cap B)$ est l'effectif de la sous-population d'individus présentant les valeurs A pour C_1 et B pour C_2.
$n(A) = n(A \cap B) + n(A \cap \overline{B})$ est l'effectif marginal (en ligne) de la valeur A de C_1.
$n(B) = n(A \cap B) + n(\overline{A} \cap B)$ est l'effectif marginal (en colonne) de la valeur B de C_2.
L'effectif total N est la somme des effectifs marginaux en ligne : $N = n(A) + n(\overline{A})$ et aussi la somme des effectifs marginaux en colonne : $N = n(B) + n(\overline{B})$.

Capacité 2 *Construire un tableau d'effectifs*

Une entreprise comprend 750 employés dont 300 cadres. De plus 60 cadres et 28 des autres employés parlent l'anglais.

Compléter le tableau croisé d'effectifs :

	Cadre C	Non cadre $\overline{C}$	Total
Parle anglais A			
Ne parle pas anglais $\overline{A}$			
Total			

1.3 Fréquences marginales

Définition 3

Étant donné un tableau croisé d'effectifs, la **fréquence marginale** d'une valeur d'un caractère est le quotient de l'effectif marginal $n(A)$ de cette valeur par l'effectif total N de la population.

$$\text{FréquenceMarginale}(A) = f(A) = \frac{\text{Effectif marginal de } A}{\text{Effectif total}} = \frac{n(A)}{N}$$

Capacité 3 *Calculer des fréquences marginales à partir d'un tableau d'effectifs*

On reprend la situation de la capacité 1.

1. Pour chacune des valeurs C et $\overline{C}$ du caractère *défaut de couleur*, puis E et $\overline{E}$ du caractère *défaut d'étanchéité*, calculer leur fréquence marginale.
2. Que vaut la somme des fréquences marginales des valeurs du caractère *défaut de couleur*? Explication?
3. Que vaut la somme de toutes les fréquences marginales en colonne?

1.4 Fréquences conditionnelles

Définition 4

Étant donné un tableau croisé d'effectifs, soit A une valeur d'un des deux caractères et C une valeur de l'autre caractère.
La **fréquence conditionnelle** de la valeur A parmi la valeur B est le quotient du nombre d'individus pré-

sentant les valeurs A et B par l'effectif marginal de la valeur B. Elle se note $f_B(A)$.

$$f_B(A) = \frac{\text{Effectif de } (A \text{ et } B)}{\text{Effectif marginal de } B} = \frac{n(A \cap B)}{n(B)}$$

Capacité 4 *Calculer des fréquences conditionnelles à partir d'un tableau d'effectifs*

On reprend la situation de la capacité 2.

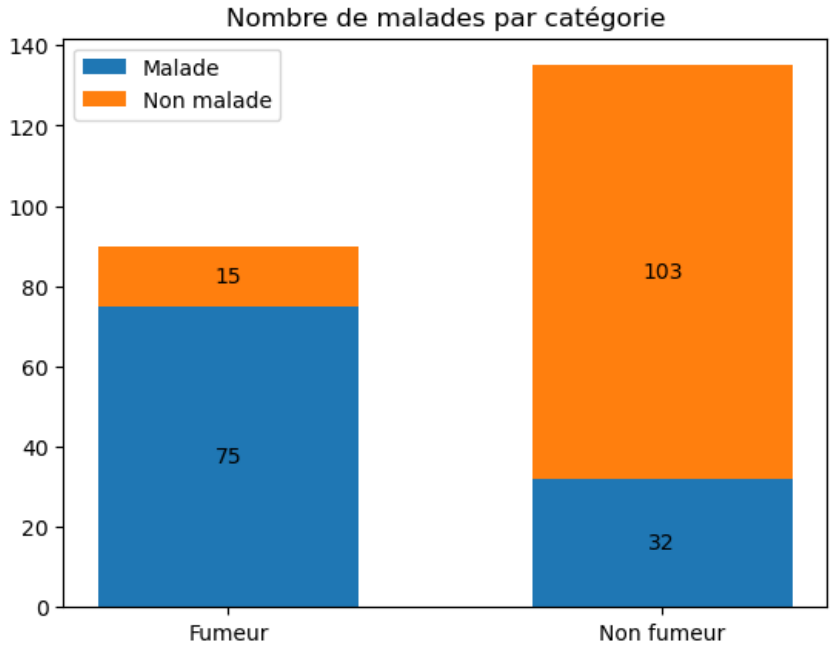
1. Calculer la fréquence des cadres qui parlent anglais parmi la population totale.
2. Calculer la fréquence conditionnelle des cadres parmi les employés qui parlent anglais. Comment la note-t-on ?
3. Que représente la fréquence conditionnelle $f_C(A)$? La calculer.
4. La capacité d'un employé à parler anglais dépend-elle du fait qu'il soit cadre ou non ? Justifier.

1.5 Représentations graphiques

Définition 5

Un **diagramme en barres** (ou « bar chart » en anglais)) est un graphique représentant une série statistique.

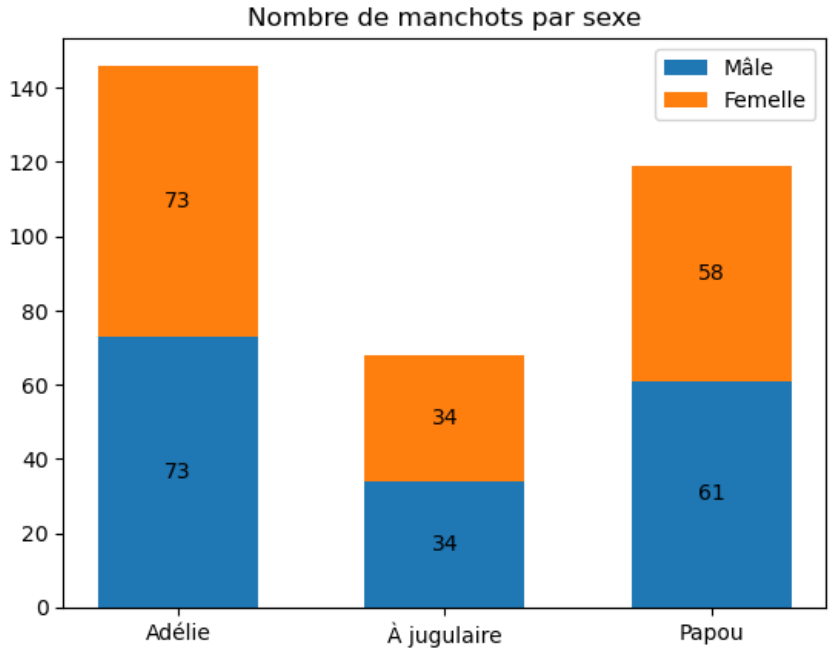
- Si on considère un seul caractère, chacune de ses valeurs est représentée par un segment ou rectangle vertical, de largeur constante et dont la longueur est proportionnelle à l'effectif (ou la fréquence).
- Si on considère deux caractères, on peut représenter un caractère avec un diagramme en barres puis découper chaque barre proportionnellement selon les valeurs du second caractère.



Une étude fait le lien entre une maladie respiratoire et le tabagisme. Le diagramme en barres ci-contre représente les effectifs des deux valeurs du caractère « Être malade ». Chaque barre est segmentée selon les deux valeurs du caractère « Être fumeur ».

Capacité 5

Les données représentées par le diagramme en barres ci-dessous ont été récoltées par le docteur Kristen Gorman à la station Palmer en Antarctique. https://allisonhorst.github.io/palmerpenguins/



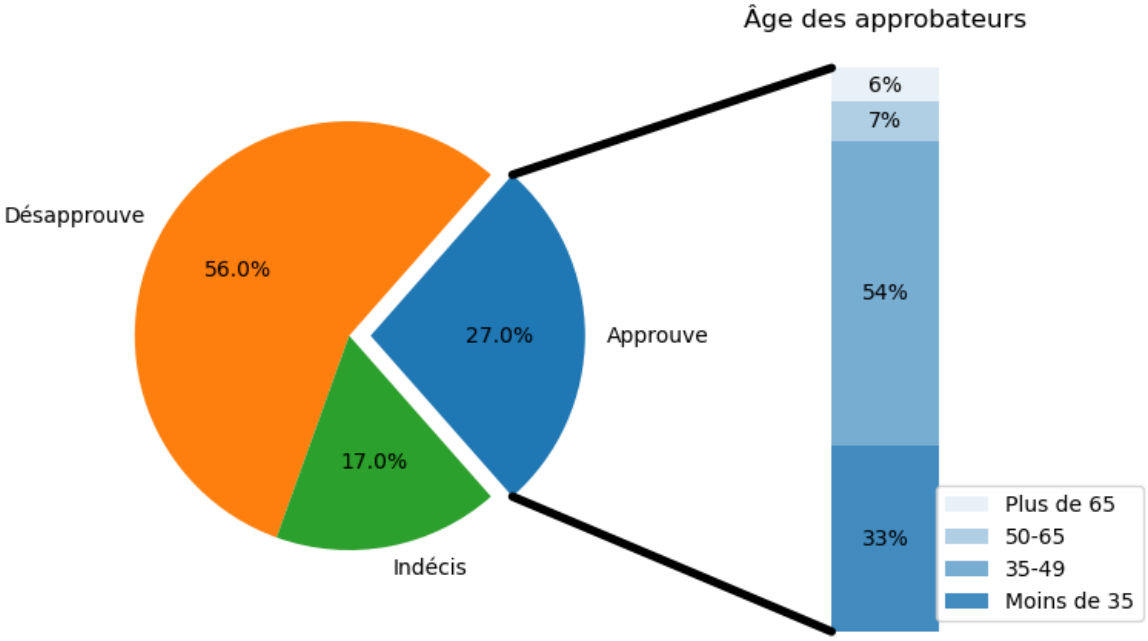
1. Quels sont les deux caractères étudiés ?
2. Représenter ces données dans un tableau croisé d'effectifs.
3. Calculer la fréquence marginale des manchots papous.
4. Calculer la fréquence marginale des manchots femelles.
5. Calculer la fréquence conditionnelle des manchots femelle parmi les manchots papous.
6. Calculer la fréquence conditionnelle des manchots papous parmi les manchots femelles.

Définition 6

Un **diagramme circulaire** (ou « pie chart » en anglais) permet de représenter les valeurs d'un caractère sous la forme d'un disque partagé en secteurs angulaires dont la mesure de chaque angle est proportionnelle à l'effectif ou fréquence de la valeur du caractère.

Capacité 6

Le diagramme circulaire ci-dessous représente les résultats d'un sondage à propos d'un projet de piétonisation des rues devant l'entrée des écoles primaires d'une commune.

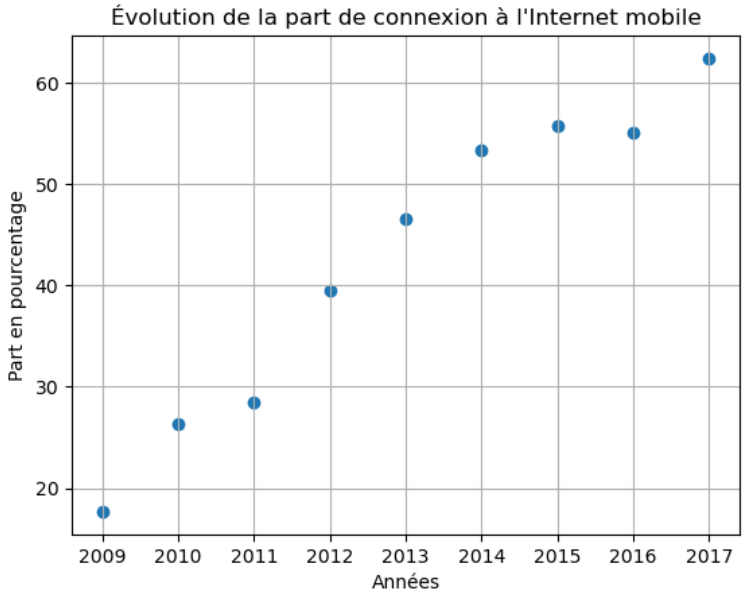


1. Déterminer la fréquence des personnes qui désapprouvent le projet par rapport à l'ensemble des personnes sondées.
2. Déterminer la fréquence conditionnelle des plus de 50 ans parmi les personnes sondées qui approuvent le projet.
3. Déterminer la fréquence des personnes sondées qui ont plus de 50 ans et qui approuvent le projet par rapport à l'ensemble des personnes sondées.

Définition 7

On considère une série statistique avec deux caractères. L'ensemble des points ayant, dans un repère donné, pour abscisse une valeur d'un caractère et pour ordonnée la valeur correspondante de l'autre caractère est appelé **nuage de points**.

Le tableau ci-dessous fournit les résultats d'une enquête de l'INSEE pour les connexions à l'Internet mobile et présente la part des personnes de plus de 15 ans résidant en France (en pourcentage arrondi au dixième) qui se sont connectées sur une période fixe.



Année	2009	2010	2011	2012	2013	2014	2015	2016	2017
Part en pourcentage : (Internet mobile)	17,7	26,4	28,4	39,5	46,5	53,4	55,8	55,1	62,4

Source : https :/ /www.insee.fr consulté le 15/01/2019

2 Probabilités conditionnelles

2.1 Des fréquences aux probabilités

Définition 8

Dans le cas d'un tirage aléatoire dans une population finie, tous les individus ont la même probabilité d'être choisis. Une sous-population réalisant la valeur d'un caractère correspond alors à un événement et sa fréquence peut être identifiée à une probabilité.

Vocabulaire des statistiques	Population	Individu	Sous-population	Fréquence
Vocabulaire des probabilités	Univers	Issue élémentaire	Événement	Probabilité

2.2 Rappels sur les lois de probabilités sur des univers finis

Définition 9

- Une loi de probabilité $\mathbb{P}$ sur l'univers fini $\Omega = \{e_1 \ ; \ e_2 \ ; \ \cdots \ ; \ e_n\}$ d'une expérience aléatoire est une fonction de Ω dans $[0\,;1]$ qui à chaque issue élémentaire e_i associe un nombre $\mathbb{P}(\{e_i\})$ compris entre 0 et 1 telle que la somme des probabilités de toutes les issues soit égale à 1 :

$$\mathbb{P}(\{e_1\}) + \mathbb{P}(\{e_2\}) + \cdots + \mathbb{P}(\{e_n\}) = 1 \tag{1}$$

  $\mathbb{P}(\{e_i\})$ est la **probabilité** de l'issue e_i.

- Toute partie A de Ω est appelée **événement** et la probabilité de A selon la loi $\mathbb{P}$ est la somme des probabilités des issues élémentaires réalisant A.

  D'après la définition d'une loi de probabilité, pour tout événement $A \subset \Omega$ on a $0 \leqslant \mathbb{P}(A) \leqslant 1$

Théorème 1

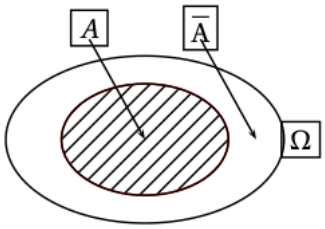
Une loi de probabilité P sur un univers fini est dite **équiprobable** (on parle aussi de loi **uniforme**) si toutes les issues ont la même probabilité. La probabilité d'un événement est alors proportionnelle au nombre d'issues élémentaires qui le constituent :

$$\mathbb{P}(\text{A}) = \frac{\text{nombre d'issues qui réalisent A}}{\text{nombre total d'issues de } \Omega} \tag{2}$$

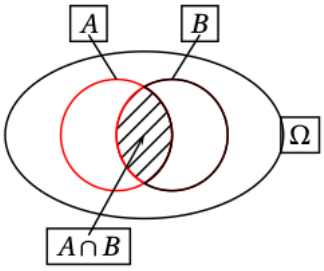
Définition 10 *Opération sur les événements*

Soit Ω l'univers d'une expérience aléatoire et A et B deux événements inclus dans Ω.

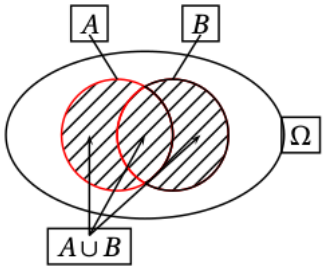
1. L'événement contraire de A dans Ω, noté $\overline{A}$ est l'ensemble des issues de Ω qui ne réalisent pas A.



2. L'**intersection** de A et de B,notée $A \cap B$, est l'ensemble des issues qui réalisent à la fois A et B.
Si $A \cap B = \emptyset$, on dit que A et B sont **incompatibles**.



3. La **réunion** de A et de B,notée $A \cup B$, est l'ensemble des issues qui réalisent A ou B (au moins l'un des deux).
On a donc $(A \cap B) \subset (A \cup B)$.



Propriété 1

Soit Ω l'univers d'une expérience aléatoire et A et B deux événements inclus dans Ω et une loi de probabilité P définie sur Ω .

- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ *Formule du crible*
- Si A et B sont incompatibles, on a $\mathbb{P}(A \cap B) = 0$ et $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$
- $\mathbb{P}\left(\overline{A}\right) = 1 - \mathbb{P}(A)$

Capacité 7 *Calculer une probabilité avec un tableau d'effectifs*

Une classe de première comporte 33 élèves. 15 pratiquent le hand-ball (noté H), 8 le tennis (noté T) et 17 ne pratiquent ni l'un ni l'autre. On choisit un élève au hasard dans cette classe.
Calculer la probabilité qu'il pratique :

- les deux sports ;
- l'un au moins des deux sports.

	H	$\overline{H}$	Total
T			8
$\overline{T}$		17	
Total	15		33

2.3 Probabilité conditionnelle

Définition 11

Soit l'univers fini d'une expérience aléatoire, muni d'une loi de probabilité $\mathbb{P}$. Soit A et B deux événements tels que $\mathbb{P}(B) \neq 0$.
La **probabilité conditionnelle** de l'événement A sachant que l'événement B s'est réalisé, notée $\mathbb{P}_B(A)$ est définie par :

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

En général $\mathbb{P}_B(A) \neq \mathbb{P}_A(B)$ *et* $\mathbb{P}_B(A) \neq \mathbb{P}(A \cap B)$.

Méthode

Si on dispose d'un tableau croisé d'effectifs par rapport à deux caractères étudiés sur une population, la probabilité conditionnelle $\mathbb{P}_B(A)$ est simplement la fréquence conditionnelle $f_B(A) = \dfrac{\text{Effectif de } (A \text{ et } B)}{\text{Effectif marginal de } B}$.

Capacité 8 *Calculer des probabilités conditionnelles avec un tableau croisé d'effectifs*

Un nouveau logiciel permet de filtrer les messages sur une messagerie électronique.
Les concepteurs l'ont testé pour 1 000 messages et voici leurs conclusions :

- 70 % des messages entrants sont indésirables ;
- 95 % des messages indésirables sont éliminés ;
- 2 % des messages bienvenus sont éliminés.

On considère les événements B : « le message est bienvenu », I : « le message est indésirable », E : « le message est éliminé » et C : « le message est conservé ».

1. Compléter le tableau suivant :

	Nombre de messages indésirables	Nombre de messages bienvenus	Total
Nombre de messages éliminés			
Nombre de messages conservés			
Total			1 000

2. Un message est reçu. Utiliser le tableau précédent pour calculer les probabilités demandées ci-dessous. Les résultats seront donnés à 10^{-3} près.

a. Exprimer en français les probabilités $\mathbb{P}_C(B)$ et $\mathbb{P}(B \cap C)$.

b. Déterminer la probabilité conditionnelle de E sachant I.

c. Déterminer les probabilités $\mathbb{P}(B \cap E)$ et $\mathbb{P}(E \cap I)$.

d. Calculer la probabilité pour que le message soit indésirable sachant qu'il est éliminé.

Capacité 9 *Application des probabilités conditionnelles aux tests médicaux*

Pour établir un diagnostic, un médecin peut utiliser un test. Ce dernier n'étant jamais parfait, il faut prendre en compte dans l'interprétation les erreurs possibles :

- un patient peut avoir un test positif sans être malade, on parle de **faux positif**;
- un patient peut avoir un test négatif en étant malade, on parle de **faux négatif**.

Partie A : sensibilité et spécificité d'un test diagnostique

Les propriétés intrinsèques d'un test sont sa **sensibilité** et sa **spécificité** :

- la **sensibilité** est la probabilité que le test soit positif sachant que le patient est malade;
- la **spécificité** est la probabilité que le test soit négatif sachant que le patient n'est pas malade.

Ces propriétés sont calculées sur un échantillon de patients dont on connaît l'état (malade ou non) à l'aide d'un autre test considéré comme sûr.

1. Quelles seraient la sensibilité et la spécificité d'un test parfait ?

2. Afin d'établir les propriétés d'un nouveau test, un laboratoire fait subir ce test à 800 malades et 800 non malades. Les résultats sont donnés dans le tableau ci-contre.

	Malade	Non malade	Total
Test positif	672	160	832
Test négatif	128	640	768
Total	800	800	1600

a. Vérifier que la sensibilité de ce test est égale à 0,84.

b. En déduire la probabilité qu'un patient malade ait un test négatif (taux de faux négatifs).

c. Calculer la spécificité de ce test.

d. En déduire la probabilité qu'un patient non malade ait un test positif (taux de faux positifs).

Partie B : valeur prédictive d'un test diagnostique

Dans cette partie, on suppose que la sensibilité d'un test est de 0,9 et que sa spécificité est de 0,8. Ce test est réalisé sur 1000 patients. On fait l'hypothèse que 75 % des patients sont malades.

1. Construire un tableau similaire à celui de la partie A.

2. Un patient a un test positif. Déterminer la probabilité qu'il soit malade.

Cette probabilité est la **valeur prédictive positive** notée **VPP**.

3. Un patient a un test négatif. Déterminer la probabilité qu'il ne soit pas malade.

Cette probabilité est la **valeur prédictive négative** notée **VPN**.