# Chapitre III

# Application de Biostatique en Epidémiologie

# Récapitulation des données

Les données existent sous forme de variables numériques ou catégorielles.

• Les variables numériques comprennent les numérations, par exemple le nombre d’enfants d'un âge donne et les mesures, par exemple la taille et le poids.

• Les variables catégorielles sont le résultat de la classification. Par exemple, on peut classer des sujets en différentes catégories en fonction de leur groupe sanguin : A, B, O ou AB. Les données ordinales – qui expriment un ordre – sont un type de données catégorielles.

# Tableaux et graphiques

## Avantages des graphiques

• simplicité et clarté

• images visuelles mémorisables

• illustration de relations complexes.

Ils mettent également l'accent sur les nombres et ont tendance a être apprécies, comme le montre leur utilisation dans les publications générales ou les tableaux sont rares.

## Avantages des tableaux

• présentation avec précision et souplesse de données plus complexes

• préparation exigeant moins de compétences ou d'installations techniques

• utilisation de moins d'espace pour une somme d'informations données.

# Mesures récapitulatives

**La moyenne**

La valeur moyenne de l'échantillon ou moyenne est la plus importante et souvent la plus appropriée ; pour un échantillon de n valeurs d'une variable telle que xi = poids
corporel elle serait égale à :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

## La médiane

La médiane est par définition la valeur centrale une fois que toutes les mesures ont été ordonnées en fonction de leurs valeurs (elle divise la distribution en 2 parties égales).

La médiane est particulièrement utile lorsque quelques valeurs sont beaucoup plus importantes que les autres

**la variable est discrète**

Le calcul de la médiane se fait à partir des effectifs ou des fréquences cumulées. La médiane est la valeur de la variable à laquelle est associé un effectif cumulé égal à N / 2, ou une fréquence cumulée égale à 0,5, **N étant effectif total de la population**.

**Exemple :** soit la distribution statistique d'une population de 30 élèves d'une classe selon leur âge :

| Âge $x_i$ | Effectifs $n_i$ | Effectifs cumulés $N_i$ |
|---|---|---|
| 14 | 6 | 6 |
| 15 | 10 | 16 |
| 16 | 10 | 26 |
| 17 | 2 | 28 |
| 18 | 2 | 30 |
| ∑ | 30 | 30 |

On a : N = 30 ; donc : N/2 = 15.

## La variable est continue

Le calcul de la médiane se fait alors en deux temps :

**Détermination de la classe médiane :** la classe médiane est **la classe de valeurs de la variable contenant la médiane.** Elle est déterminée de la même manière que la médiane dans le cas d'une variable discrète, à partir des effectifs et des fréquences cumulés.

**Exemple :** soit la distribution statistique d'une population de 30 élèves d'une classe selon leur taille :

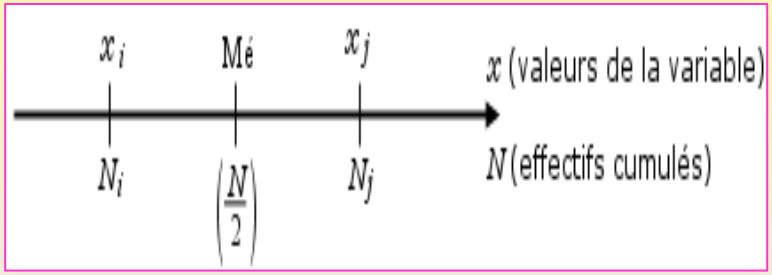
| Taille $x_i$ | Effectifs $n_i$ | Effectifs cumulés $N_i$ |
|---|---|---|
| <1,60 | 8 | 8 |
| [1,60-1,70[ | 9 | 17 |
| [1,70-1,80[ | 10 | 25 |
| [1,80-1,90[ | 2 | 27 |
| ≥ 1,90 | 1 | 30 |
| ∑ | 30 | 30 |

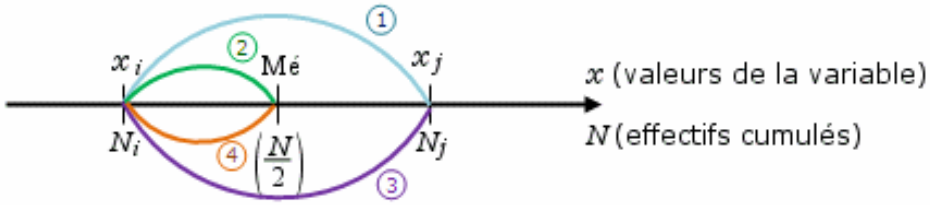
**La classe [1,60-1,70[ est la médiane de la distribution.**

**Détermination de la médiane :** cette seconde étape cherche à découvrir la valeur précise de la médiane à l'intérieur de la classe médiane.

La méthode généralement utilisée pour ce faire est celle de l'interpolation linéaire ; c'est mathématiquement une application simple du théorème de Thalès.

Soit $[x_i; x_j[$ la classe médiane déterminée à l'étape précédente. $N_i$ et $N_j$ les effectifs cumulés associés aux deux bornes de cette médiane : $x_i$ et $x_j$




$$\frac{\text{Segment ②}}{\text{Segment ①}} = \frac{\text{Segment ④}}{\text{Segment ③}}$$

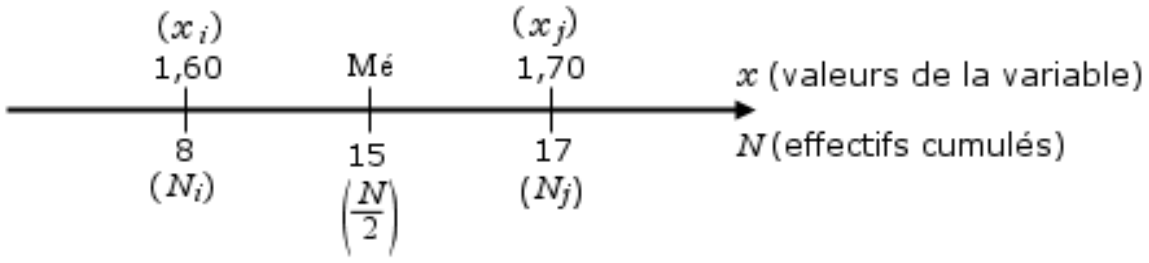
Soit :

$$\frac{(Mé - x_j)}{(x_j - x_i)} = \frac{\left(\frac{N}{2} - N_i\right)}{(N_j - N_i)}$$

D'où :

$$Mé = x_i + \frac{\left(\frac{N}{2} - N_i\right) \times (x_j - x_i)}{(N_j - N_i)}$$

**Exemple :** dans l'exemple précédent, on a : $x_i$ = 1,60 m et $x_j$ = 1,70 m (bornes de la classe médiane) ; on a aussi : $N_i$ = 8 et $N_j$ = 17 (effectifs cumulés associés) ;
on peut alors construire l'axe suivant, soit :




Soit :

$$\frac{(\text{Mé} - 1{,}60)}{(1{,}70 - 1{,}60)} = \frac{15 - 8}{17 - 8}$$

D'où :
**Mé = 1,68 m**

## Le mode

Le mode est **la valeur de la variable la plus fréquente de la population étudiée**.
En d'autres termes, dans une distribution statistique, le mode est la modalité de la variable à laquelle est associé le plus grand effectif ou la plus grande fréquence. On note généralement le mode : $M_0$.

Le calcul du mode de distribution et sa difficulté dépendent de la nature continue ou discrète de la variable étudiée.

## Cas de la variable discrète

Le mode est la valeur de la variable possédant le plus grand effectif ou la plus grande fréquence. Il est, dans ce cas, simplement ou directement observable. Dans un tableau statistique, c'est le $x_i$ ou le $f_i$ le plus élevé.

**Exemple :**

soit la distribution statistique d'une population de 30 élèves d'une classe selon leur âge, dont le tableau statistique est :

| Âge $x_i$ | Effectifs $n_i$ | Fréquences $f_i$ | Fréquences $f_i$ en % |
|---|---|---|---|
| 14 | 6 | 0,2 | 20 |
| **15** | **10** | 0,33 | 53 |
| **16** | **10** | 0,33 | 86 |
| 17 | 2 | 0,067 | 93 |
| 18 | 2 | 0,067 | 100 |
| ∑ | 30 | 1 | 100 % |

L'effectif $n_i$ ou la fréquence $f_i$ les plus élevés montrent que le mode est ici de 15 et 16 ans (l'effectif est le même dans les deux cas).

## Cas de la variable continue

Si la variable est continue, ses modalités sont des **classes de valeurs**.
Le mode de distribution ne pourra pas être une modalité représentant une valeur précise de cette variable mais sera une classe de valeurs.

On appelle alors **classe modale** la classe constituant **le mode de la distribution**.

## Exemple

soit la distribution statistique d'une population de 30 élèves d'une classe selon leur taille :

| Taille $x_i$ | Effectifs $n_i$ | Fréquences $f_i$ | Fréquences $f_i$ en % |
|---|---|---|---|
| <1,60 | 8 | 0,267 | |
| [1,60-1,70[ | 9 | 0,30 | |
| **[1,70-1,80[** | **10** | 0,33 | |
| [1,80-1,90[ | 2 | 0,067 | |
| ≥ 1,90 | 1 | 0,033 | |

## Variances, écarts types et erreurs types

Les mesures de la dispersion constituent un autre groupe de mesures récapitulatives.

Les trois mesures de dispersion les plus utiles:

- La variance
- L’écart type
- L’erreur type

## La variance

La variance est, selon la définition classique, la moyenne des carrés des écarts par rapport à la moyenne. Elle peut être considérée comme une mesure servant à caractériser la dispersion d'une distribution ou d'un échantillon. La formule de la variance est la suivante :

$$S^2 = \frac{\sum (x - \bar{x})^2}{n}$$

**($S^2$) = moyenne de l'écart au carré de valeurs par rapport à la moyenne**

**Exemple**

Les notes d'un élève sont : 12,12,12,12,10,14

La variance de cette série est: V(x) = 4/3 ou la moyenne est égale à 12

## Ecart type

σ, est la racine carré de la Variance: $\sigma(X) = \sqrt{V(X)}$

**Exemple**

L'écart type de l'exemple précédent est: $\sigma \approx 1{,}15$

## Erreur type

L'erreur type de la moyenne reflète la manière dont toutes les moyennes possibles des échantillons de taille n pourraient être dissemblables si chaque échantillon était choisi au hasard dans la même population que l'échantillon initial.

L'erreur type égale à: $\boldsymbol{\sigma/\sqrt{n}}$

## Intervalles de confiance

Les intervalles de confiance sont l'un des instruments les plus utiles de l’épidémiologie.

Un intervalle de confiance se sert de ces concepts pour définir des limites raisonnables applicables à la moyenne de la population, d’après les données d’un échantillon.

Les intervalles de confiance sont faciles a préparer et relativement faciles à comprendre.

**Exemple**
On choisit 10 personnes au hasard dans une population et leurs poids sont mesures en kilogrammes : 82.3, 67.3, 68.6, 57.7, 67.3, 60.5, 61.8, 54.5, 73.2 et 85.9.

-Calculer la moyenne de ces poids.

-Construire l'intervalle de confiance

**La moyenne: M** =67,9 kg

Pour construire un intervalle de confiance, on calcule une limite inferieure et une limite supérieure.
Pour l'échantillon des poids, avec n = 10, l'écart type calcule est de σ = 10,2 kg.

Les limites inferieures et supérieures sont les suivantes :

Limite inférieure: $M - (2,6)\ \sigma / \sqrt{n} = 67,9 – 2,26(10,2) / 3,16 = 60,61$
Limite supérieure: $M + (2,6)\ \sigma / \sqrt{n} = 67,9 + 2,26(10,2) / 3,16 = 75,19$

L’intervalle de confiance qui en résulte: $C(60,61 < \mu < 75,19) = 0,95$

Il indique clairement qu’il y a un intervalle de confiance a 95 % pour la moyenne de la population.

La longueur de cet intervalle est de 75,19 – 60,61 = 14,58 kg,

On peut utiliser un intervalle de confiance pour tester une hypothèse, à savoir l’hypothèse que $\mu = 80,0$ kg.

Dans ce cas, l’hypothèse a été testée et rejetée en se basant sur les limites inferieure et supérieure de l’intervalle de confiance.
En général, les intervalles de confiance peuvent être utilises ainsi pour tester les hypothèses

## Tests d’hypothèse, valeurs de p, puissance statistique

Les tests d’hypothèse sont relativement directs.

**la valeur de p** associée a ce test et la puissance statistique qu’a le test pour ≪ déceler ≫ une différence d’une ampleur donnée.

valeur de p représente la probabilité pour qu'une valeur de la moyenne d'un échantillon aléatoire de la population étudiée soit située à la moyenne calculée ou a une valeur encore plus éloignée d'une valeur considérée comme hypothèse de chaque côté.

## Puissance statistique

Dans la description du test t a deux échantillons qui suit, il est fait référence a l'hypothèse nulle :

$H_0: M_1 - M_2 = 0$

$H_1: M_1 - M_2 \neq 0$

qui examine les différences entre les moyennes de deux populations. S'il s'agit des poids corporels de deux populations.

Plus la différence entre les moyennes des deux populations sera importante, plus il sera facile de rejeter cette hypothèse nulle a l'aide des moyennes des échantillons.

# Méthodes statistiques de base en épidémiologie

- **Test t**

En épidémiologie, il est fréquent d'avoir deux échantillons représentant deux populations différentes et de vouloir répondre a la question de savoir si les moyennes des deux échantillons sont suffisamment différentes pour en conclure que les deux populations qu'elles représentent ont des moyennes différentes.

Le test t (ou test de Student) a recours a une statistique qui, dans le cadre de l'hypothèse nulle, teste si ces deux moyennes présentent une différence significative.
L'hypothèse :

$H_o: M_1 - M_2 = 0$

$H_1: M_1 - M_2 \neq 0$

est testée avec l'aide de la statistique t avec $(n_1 + n_2 - 2)$ degrés de liberté.

# Test de Student pour échantillon unique

Il s'agit de comparer une **moyenne observée** à une **moyenne théorique** ($\mu$).

Soit X une série de valeurs de taille n, de **moyenne** m et d'**écart-type** S. La comparaison de la **moyenne observée** (m) à une **valeur théorique** $\mu$ est donnée par la formule :

$$t = m - \mu \,/\, s/\sqrt{n}$$

Pour savoir si la différence est significative, il faut tout d'abord lire dans la **table**, la valeur critique correspondant au **risque alpha** = 5% pour un degré de liberté :

$d.d.l = n - 1$

Si la valeur absolue de t (|t|) est supérieure à la valeur critique, alors la différence est significative. Dans le cas contraire, elle, ne l'est pas. Le **degré de siginificativité** (ou **p-value**) correspond au risque indiqué par la **table de Student** pour la valeur |t|

## Test t de Student pour échantillons indépendants

Il s'agit de **comparer deux moyennes observées**.
Lorsque les deux groupes d'échantillons (A et B) à comparer n'ont aucun lien, on utilise le **test t de Student indépendant** (ou **non apparié**).

### Formule

$$t = mA - mB \,/\, \sqrt{S^2/n_A + S^2/n_B}$$

$S^2$ est la **variance** commune aux deux groupes. Elle est calculée par la formule suivante :

$$S^2 = \sum (x - mA)^2 + (x - mB)^2 \,/\, nA + nB - 2$$

Si la valeur absolue de t (|t|) est supérieure à la valeur critique, alors la différence est significative. Dans le cas contraire, elle, ne l'est pas.

Le **degré de siginificativité** ou **p-value** correspond au risque indiqué par la **table de Student** pour la valeur |t|

**Exemple**

Un groupe de 100 individus (50 femmes et 50 hommes) pris au hasard au sein de la population.

On se pose la question à savoir si le poids moyen des femmes est significativement différent de celui des hommes?

Dans cet exemple on parle de **test de Student non apparié** car les deux groupes à comparer n'ont aucun lien.

Il s’agit donc de calculer le poids moyen des femmes et de celui des hommes et d’évaluer si la **différence est significative** au point de vue **statistique**.

## Test-t de Student pour séries appariés

Le **test de Student apparié** permet de comparer la moyenne de deux séries de valeurs ayant un lien.

**Exemple**

20 souris ont reçu un traitement X pendant 3 mois. On se pose la question à savoir si le traitement X a un impact sur le poids des souris au bout des 3 mois. Le poids des 20 souris a donc été mesuré avant et après traitement. Ce qui nous donne 20 séries de valeurs avant traitement et 20 autres séries de valeurs après traitement provenant de la mesure du poids des mêmes souris.

La **valeur t de Student** est donnée par la formule :

$$t = m / s/\sqrt{n}$$

**m** et **s** représentent la **moyenne** et l'**écart-type** de la différence d  des mesures entre les paires des valeurs.
**n** est la taille de la série d.

# Test du Khi-carré pour les tableaux à entrées multiples

Les tableaux a entrée multiple, ou tableaux de contingence, sont des instruments qui permettent de présenter le nombre de participants classes en fonction de deux ou plusieurs facteurs ou variables.

Le tableau suivant est un exemple typique avec r = 2 rangées et c = 2 colonnes de données pour un tableau r x c ou 2 x 2.

Ce tableau présente l'association entre deux catégories d'exposition et deux catégories d'état morbide.

**Tableau: Association entre la consommation de viande et l'entérite nécrosante**

| | Exposition (ingestion récente de viande) | | |
|---|---|---|---|
| | oui | non | total |
| Présence de maladie | 50 | 11 | 61 |
| Absence de maladie | 16 | 41 | 57 |
| Total | 66 | 52 | 118 |

Tester l'hypothèse :

**H0** : Il n'y a pas d'association entre cette classification de l'exposition et cette classification de l'état morbide , ou

**H1** : Il y a association entre cette classification de l'exposition et cette classification de l'état morbide.

Pour les tableaux 2 x 2, cette hypothèse peut également porter sur les comparaisons entre deux proportions:
PE = Proportion des personnes exposées qui ont présente la maladie,
PNE = Proportion des personnes non exposées qui ont présente la maladie, de sorte que l'hypothèse peut être exprimée comme suit :

$H_0 : P_E = P_{NE}$, ou $H_1 : P_E \neq P_{NE}$.

Pour tester cette hypothèse on compare la fréquence observée, O dans chaque case à la fréquence escomptée, E, que l'on aurait si l'hypothèse nulle était complètement
vraie. On peut calculer E afin de créer le tableau suivant :

**E=Total de la rangée contenant la cellule) x (Total d e la colonne contenant la cellule) / Total général du tableau**

| Cellule | O | E | O-E | $(O-E)^2$ | $\frac{(O-E)^2}{E}$ |
|---|---|---|---|---|---|
| 1 | 50 | 34,12 | 15,88 | 252,22 | 7,39 |
| 2 | 11 | 26,88 | -15,88 | 252,22 | 9,38 |
| 3 | 16 | 31,88 | -15,88 | 252,22 | 7,91 |
| 4 | 41 | 25,12 | 15,88 | 252,22 | 10,04 |
| Total | 118 | 118 | 0,00 | | **34,72** |

Le total de la dernière colonne est la valeur calculée pour $\chi^2$ qui correspond au résultat du test de Khi-carre avec 1 degré de liberté. En général, le nombre de degré de liberté est de df = (r–1)x(c–1).

La valeur calculée, à savoir 34,72, est bien plus élevée que celle figurant dans le tableau du Khi-carre pour **α** = 0,05, qui est de 3,84 ; par conséquent, l'hypothèse nulle est rejetée.

**NB/**

Des tables de distribution du Khi-carre sont disponibles en ligne ou dans n'importe quel manuel de statistiques classique.

# Corrélation

En général, la corrélation quantifie le degré de variation de deux variables entre elles

Si ces deux variables sont indépendantes alors la valeur de l'une n'a aucun rapport avec la valeur de l'autre.

Si elles sont corrélées; la valeur de l'une est liée a la valeur de l'autre, c'est-a-dire que l'une est élevée lorsque l'autre est élevée ou que l'une est élevée lorsque l'autre est faible. Il existe plusieurs instruments pour mesurer la corrélation.

Le plus communément employé est le coefficient de corrélation des moments mixtes, dit coefficient de corrélation de Pearson qui se calcule au moyen de l'équation :

$$r = \frac{COV(x, y)}{S_x S_y}$$

la covariance est calculée par l'équation:

$$COV(x, y) = \frac{\sum (x_i - \overline{X})(y_i - \overline{Y})}{n - 1}$$

Ce coefficient mesure le degré de corrélation linéaire et est situé entre $-1 \leq r \leq 1$.

Il est proche de +1 lorsqu'il y a une forte association linéaire positive et est proche de -1

lorsqu'il y a une forte association négative, c'est-a-dire lorsqu'une faible valeur de x a  tendance à impliquer une forte valeur de y.

Lorsque r = 0, il n'y a pas d'association linéaire.

# Régression

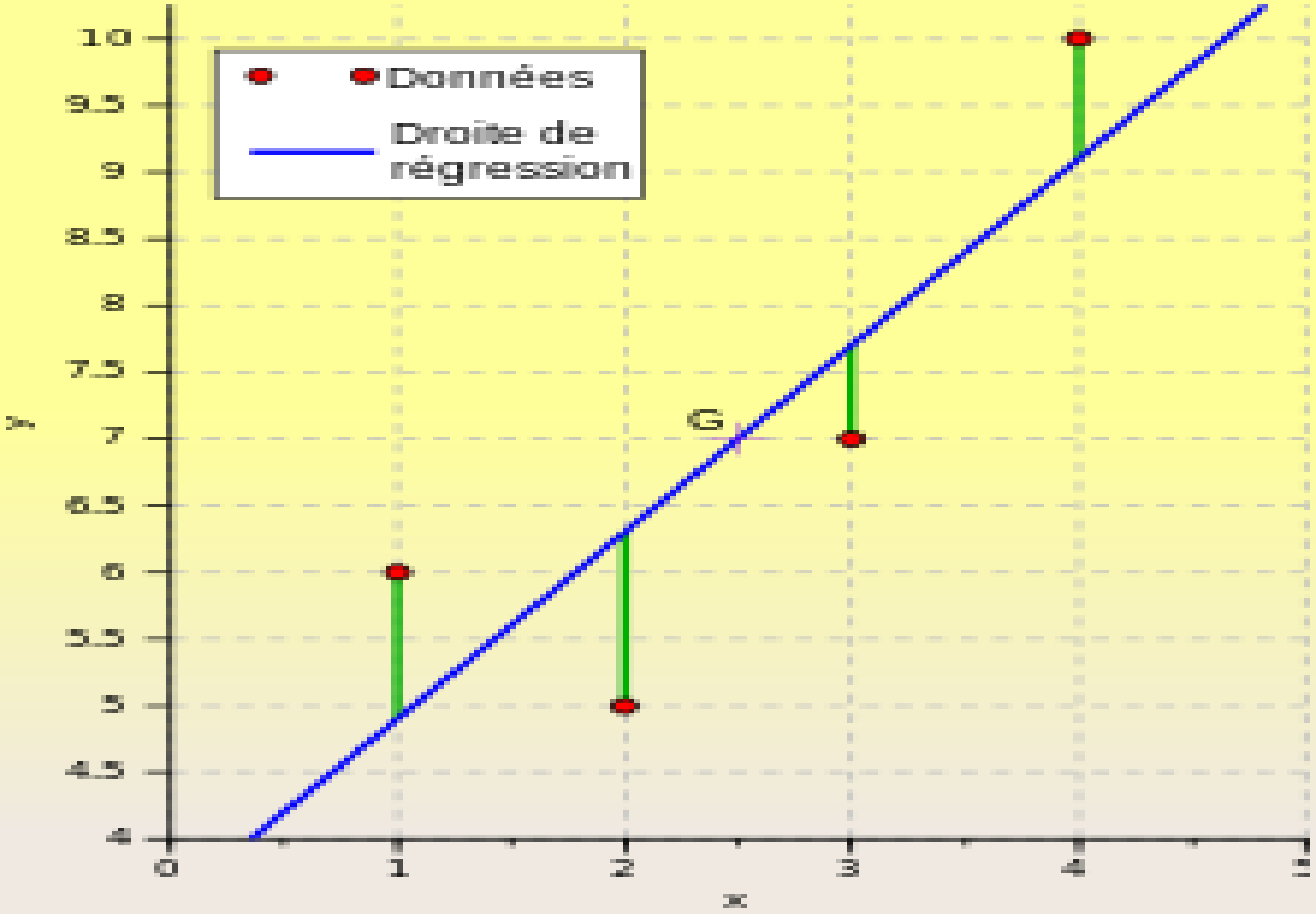
## Différents modèles de régression

Trois types de modèles de régression sont fondamentaux pour la recherche épidémiologique :

- La régression linéaire
- La régression logistique
- La régression a risques proportionnels de Cox, un type d'analyse de la survie

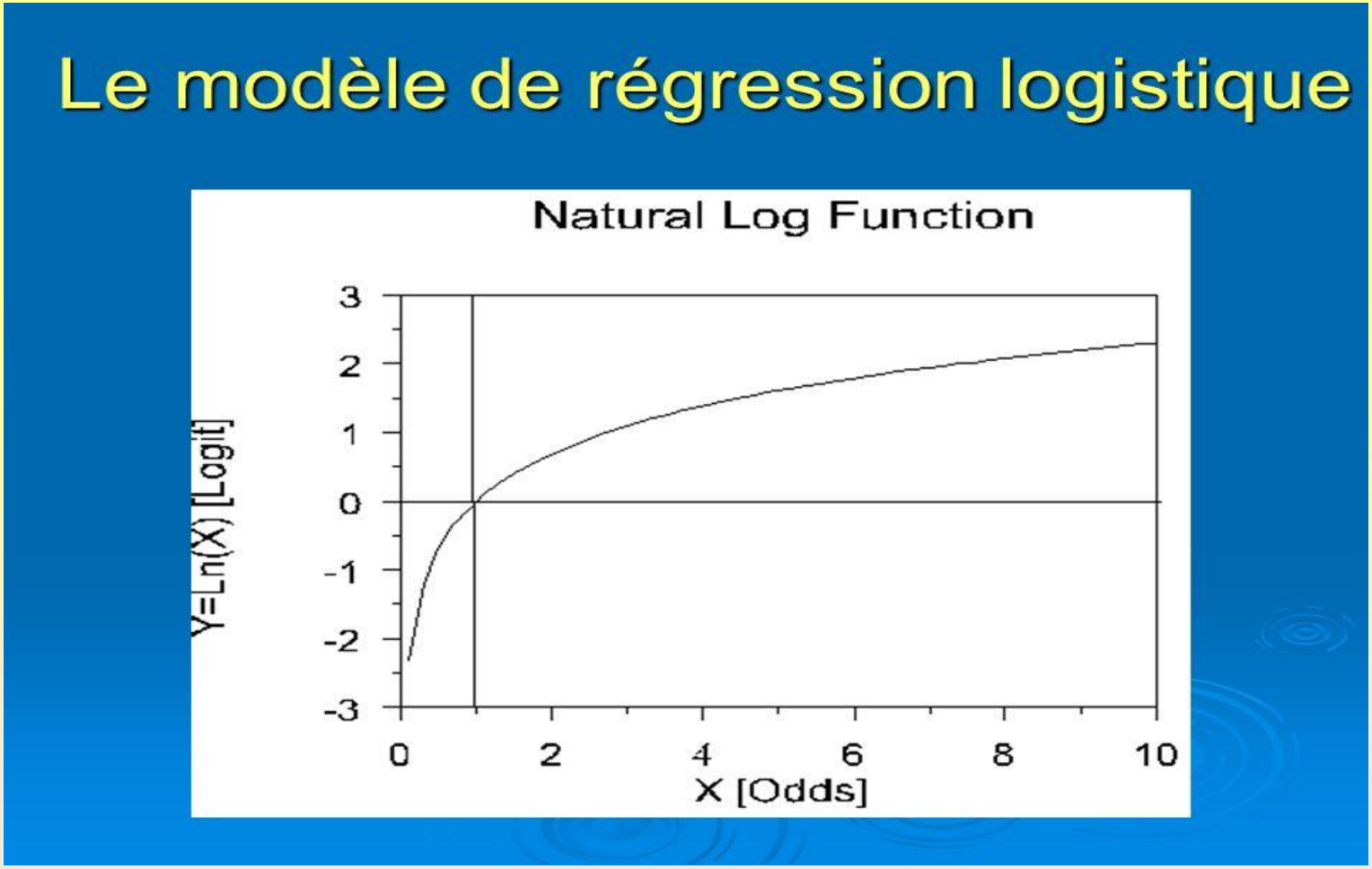
## Modèles de régression linéaire

La variable dépendante doit être une variable continue dont la distribution de fréquence est une distribution normale.


Données
Droite de régression
G
y
x
10
9.5
9
8.5
8
7.5
7
6.5
6
5.5
5
4.5
4
0
1
2
3
4
5

- **Modèles de régression logistique**

La variable dépendante est dérivée de la présence ou de l'absence d'une caractéristique, en général représentée par 0 ou 1.



• Modèles des risques proportionnels de Cox

La variable dépendante représente la durée écoulée entre un point de départ donné et la survenue d'un évènement étudié.

L'avantage d'un modèle comme celui de Cox est que l'on peut prendre en compte ces données même si elles ne sont pas « complètes ».

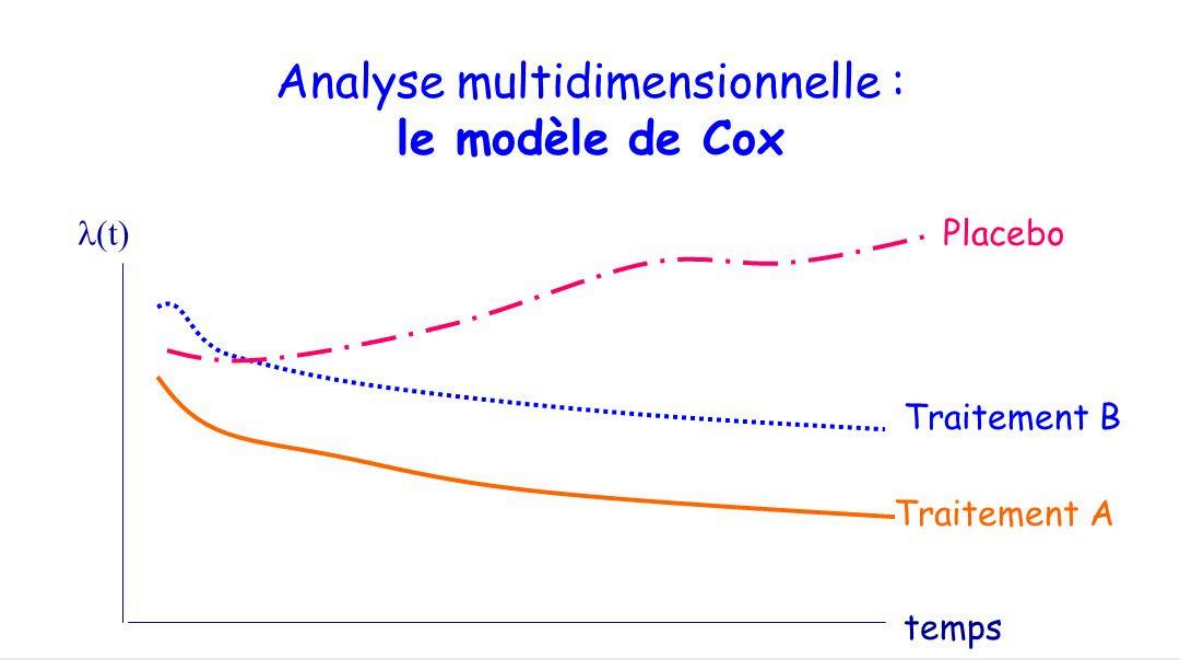
**Exemple**

Prenons une covariable qui peut prendre deux valeurs : 0 si l'individu prend le traitement A ou 1 s'il s'agit du traitement B.

Prenons comme référence les individus qui prennent le traitement A (la manière de procéder pour le codage des variables est parfaitement identique aux modèles tels que la régression logistique ou linéaire),

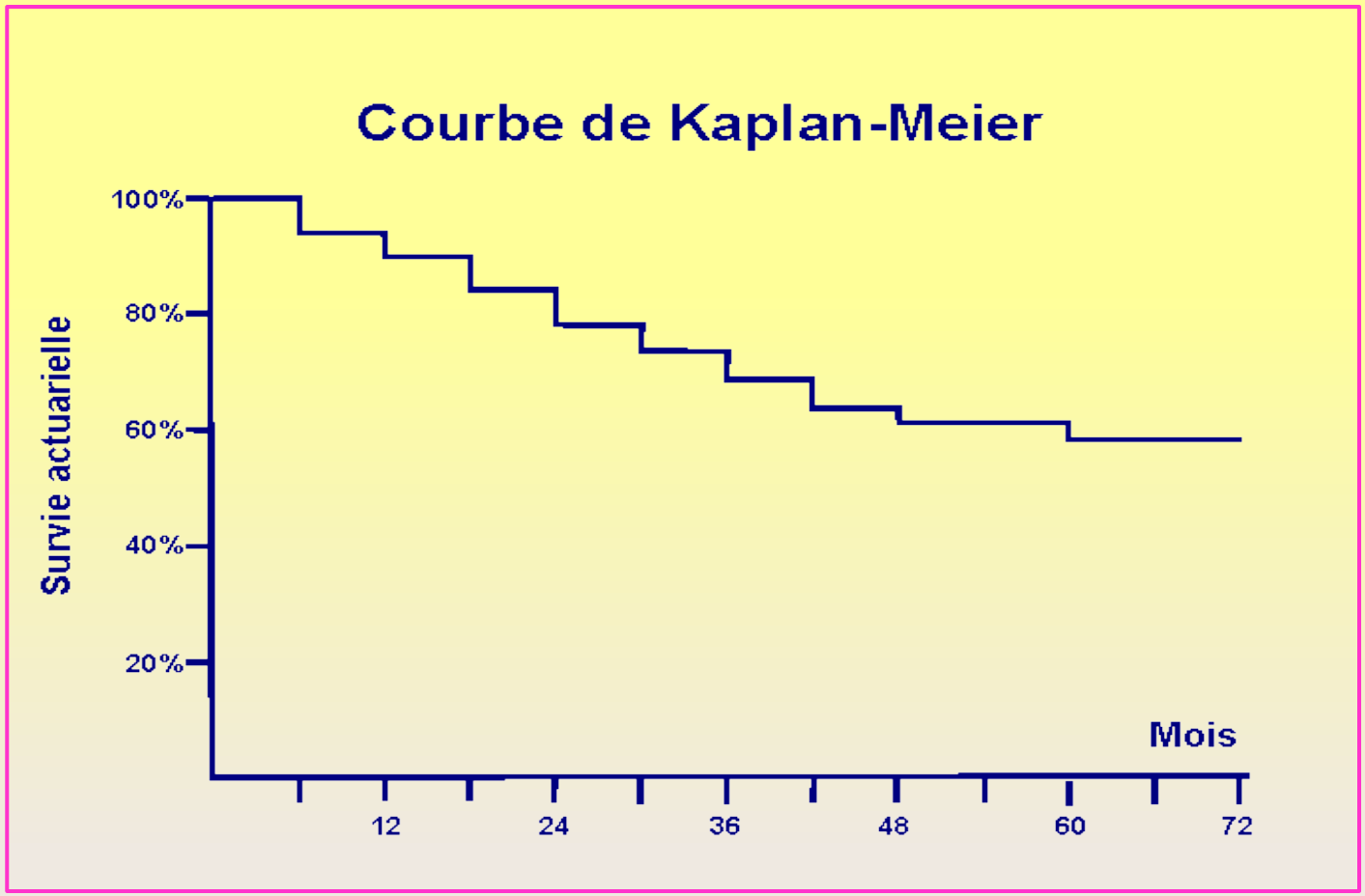
Le coefficient  est le **Risque Relatif** (ici risque instantané de décès) associé au traitement B par rapport au traitement A.

Autrement dit, les individus prenant le traitement B ont un risque instantané de décès  fois celui des individus prenant le traitement A.



# Courbes de survie de Kaplan-Meier

Ces courbes  sont couramment utilisées pour présenter les données de la survie



## Problèmes liés à la taille de l'échantillon

L'un des problèmes que l'on rencontre souvent dans les investigations épidémiologiques  est d'imaginer quelle est la taille de l'échantillon dont on a besoin pour répondre a une question particulière.

L'échantillon doit être suffisamment grand pour que l'étude ait une puissance statistique appropriée – à savoir qu'elle puisse démontrer une association s'il y en a une

On base les calculs de la taille de l'échantillon sur un certain nombre de facteurs qui interviennent dans la conception de l'étude :

- la prévalence
- la marge d'erreur acceptable
- la différence décelable (ou le degré de confiance vise).

## Méta-analyse

Par définition, la méta-analyse est la synthèse statistique des données d'études séparées mais analogues (comparables), conduisant a un récapitulatif quantifiable des résultats groupes en vue d'identifier une tendance générale.

La méta-analyse diffère de la plupart des études médicales et épidémiologiques en ce qu'aucune nouvelle donnée n'est recueillie, mais qu'on combine plutôt les résultats d'études antérieures.

Les différentes étapes pour mener une bonne méta-analyse:

- Formuler le problème et la structure de l'étude ;

- Recenser les études pertinentes ;

- Exclure les études mal conduites ou celles qui présentent des défauts méthodologiques majeurs ; et

- Mesurer, combiner et interpréter les résultats

## Un exemple

méta-analyse est celle effectuée à propos de l'utilité de manger une certaine quantité de fruits et de légumes par jour.

Cette méta-analyse a porté sur plus de 2580 000 personnes suivies pendant 13 ans.
Chez ces individus on a recherché un lien de cause à effet entre la consommation de fruits et de légumes et la survenue d'accidents vasculaires cérébraux.

Les conclusions ont été les suivantes : les individus consommant moins de trois portions de fruits et de légumes (77 g de légumes et 80 g de fruits) présentent 11 % de risque d'accident vasculaire cérébral.

Si la consommation excède les 5 portions, le risque est diminué de 26 %.
Ceci concerne autant les accidents vasculaires cérébraux de nature hémorragique (par hémorragie) qu'ischémique c'est-à-dire dus à la présence d'un caillot sanguin.