Université de Liège

Faculté de Médecine

Méthodes statistiques en épidémiologie

Pr. A. ALBERT

Théorie et Exercices

Edition 2008

Table des matières

Chapitre 1

Introduction générale

Epidémiologie
L'épidémiologie est la science qui permet:

1. d'étudier la fréquence (prévalence) des maladies dans diverses populations
2. d'en suivre l'évolution
3. de former des hypothèses sur l'étiologie (=étude des causes) et la prévention de ces maladies

Statistique
La statistique est la science qui étudie:

1. les méthodes de réduction de données (=statistique descriptive)
2. la variabilité
3. les populations (=statistique inférentielle: tests d'hypothèses, estimations)

Population
Une population est un ensemble d'individus ou d'objets qui ont au moins une propriété en commun. Il est nécessaire de bien définir la population à laquelle on s'intéresse. On note N l'effectif (size) de la population, c'est-à-dire le nombre d'individus ou d'objets qui la constitue. En général, N est tellement grand qu'il est assimilable à l'**infini**. Toutefois, en épidémiologie, il arrive que N doive être

considéré comme fini.

Echantillon
Un échantillon (*sample*) est un sous-ensemble de la population. Il doit être représentatif de la population. On note n l'effectif de l'échantillon. Il est toujours **fini**.

Variable
Une variable est une grandeur à laquelle on s'intéresse dans la population. On la note X. Il y a deux types de variables:

1. *quantitative* ou mesurée (âge, cholestérol, pression artérielle systolique, indice de masse corporelle)
2. *qualitative* ou observée (tabagisme, maladie, race, groupe sanguin)

Moyenne dans la population (*mean*)
La moyenne de X dans la population est notée μ. En général, cette quantité est inconnue. Soient $\{x_1, \cdots, x_N\}$ les valeurs de X chez les N sujets de la population.

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + \cdots + x_N}{N}$$

Pour les *variables discrètes* (celles qui prennent un nombre fini m de valeurs), la formule précédente peut aussi s'écrire:

$$\mu = \frac{\sum_{i=1}^{m} R_i X_i}{N}$$

où $X_1, \cdots, X_m$ sont les m valeurs distinctes de X et $R_1, \cdots, R_m$, leurs répétitions ($N = R_1 + \cdots + R_m$).
En particulier, pour les *variables binaires* (X=0 ou 1), la formule devient ($m = 2$):

$$\mu = \frac{(N - R) \times 0 + R \times 1}{N} = \frac{R}{N} = \pi$$

où R est le nombre de sujets pour lesquels $X = 1$ et $N - R$, le nombre de sujets pour lesquels $X = 0$.
La moyenne d'une variable binaire dans une population est une *proportion* que l'on note π.

Moyenne dans l'échantillon

La moyenne de X dans l'échantillon est notée $\overline{x}$. Cette quantité est connue. Soient $\{x_1, \cdots, x_n\}$ les valeurs de X chez les n sujets de l'échantillon.

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + \cdots + x_n}{n}$$

Pour les *variables discrètes* (celles qui prennent un nombre fini de valeurs).

$$\overline{x} = \frac{\sum_{i=1}^{m} r_i X_i}{n}$$

où $X_1, \cdots, X_m$ sont les m valeurs distinctes de X et $r_1, \cdots, r_m$ leurs répétitions dans l'échantillon ($n = r_1 + \cdots + r_m$).
En particulier, pour les *variables binaires* (X=0 ou 1)

$$\overline{x} = \frac{(n-r) \times 0 + r \times 1}{n} = \frac{r}{n} = p$$

où r est le nombre de sujets dans l'échantillon qui ont $X = 1$ et $n - r$ le nombre de sujets pour lesquels $X = 0$.
La moyenne d'une variable binaire dans un échantillon est une *proportion* que l'on note p.

Epreuves d'hypothèses

Les hypothèses se font sur les paramètres de la population (μ ou π). On teste une hypothèse nulle notée H_0 versus une hypothèse alternative notée H_1.

H_0: $\mu_1 = \mu_2$ vs H_1: $\mu_1 \neq \mu_2$: test d'égalité de moyennes

H_0: $\pi_1 = \pi_2$ vs H_1: $\pi_1 \neq \pi_2$: test d'égalité de proportions

Pour tester des hypothèses, on se base sur des données. Celles-ci sont obtenues à partir d'enquêtes comme par exemple celle du HIS (*Health Interview Survey*) qui contient des informations de santé sur 10,000 personnes en Belgique.
En général, on ne dispose que d'un échantillon de la population pour décider entre H_0 et H_1. Il y a donc un certain degré d'incertitude. Il y a deux types d'erreur possibles dans la décision finale:

- α est la probabilité de rejeter H_0 alors qu'elle est vraie.
  C'est le *risque de première espèce*.

- β est la probabilité de ne pas rejeter H_0 alors qu'elle est fausse.
  C'est le *risque de deuxième espèce*.

Pour n fixé et α fixé (souvent 0.05), β est fixé.

Pour n fixé, si α augmente (diminue), alors β diminue (augmente).

Pour diminuer α et β simultanément, il faut augmenter n.

La quantité $1 - \beta$ est la probabilité de rejeter H_0 lorsque H_0 est fausse. C'est la *puissance du test* (*power*).

Variance dans la population
Soit X la variable à laquelle on s'intéresse et μ sa moyenne dans la population. La variance est un indicateur de dispersion. La variance est toujours supérieure ou égale à zéro. La variance dans la population se note σ^2.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = \frac{\sum_{i=1}^{N} x_i^2 - N\mu^2}{N}$$

Pour les variables discrètes,

$$\sigma^2 = \frac{\sum_{i=1}^{m} R_i X_i^2 - N\mu^2}{N}$$

Pour les variables binaires,

$$\sigma^2 = \frac{(N - R) \times 0^2 + R \times 1^2 - N\pi^2}{N} = \pi \times (1 - \pi)$$

Ecart-type dans la population (*standard deviation*)
L'écart-type σ est la racine carrée positive de la variance. Il s'exprime dans les mêmes unités que la variable X: $\sigma = \sqrt{\sigma^2}$ ou $\sigma = \sqrt{\pi \times (1 - \pi)}$ pour une proportion.

Variance dans l'échantillon
Soit $\{x_1, \cdots, x_n\}$ les valeurs de X pour les n sujets de l'échantillon et $\overline{x}$ sa moyenne dans l'échantillon. La variance dans l'échantillon se note s^2.

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

Pour les variables binaires,

$$s^2 = \frac{(n-r) \times 0^2 + r \times 1^2 - np^2}{n} = p \times (1-p)$$

Ecart-type dans l'échantillon
Il s'exprime dans les mêmes unités que la variable X: $s = \sqrt{s^2}$ ou $s = \sqrt{p \times (1-p)}$ pour une proportion.

Variabilité d'échantillonnage (*sampling variability*)
La variabilité d'échantillonnage est la variabilité d'échantillon à échantillon. C'est un concept fondamental pour estimer la précision des résultats statistiques obtenus. Si l'on considère une population d'effectif N, le nombre d'échantillons d'effectif n que l'on peut extraire de cette population est considérable. Ce qui nous intéresse est de savoir si la moyenne de la variable X varie beaucoup ou peu entre les différents échantillons.
Soit un échantillon $x_1, \cdots, x_n$ extrait de la population. La moyenne vaut $\overline{x}$.
Soit un échantillon $x'_1, \cdots, x'_n$ extrait de la population. La moyenne vaut $\overline{x}'$.
Soit un échantillon $x''_1, \cdots, x''_n$ extrait de la population. La moyenne vaut $\overline{x}''$.
etc...
Y a-t-il une grande variabilité entre $\overline{x}$,$\overline{x}'$,$\overline{x}''$,...?

On peut montrer que la variance des $\overline{x}$, notée $\sigma^2(\overline{x})$, vaut:

$$\sigma^2(\overline{x}) = \frac{\sigma^2}{n} \frac{N-n}{N-1}$$

où σ^2 est la variance de X dans la population.

Le facteur $\frac{N-n}{N-1}$ est un *facteur de correction* pour les populations finies. On remarque que si la population est infinie, le facteur de correction vaut 1 et on retrouve ainsi la variance de la moyenne d'échantillon habituelle, c'est-à-dire $\sigma^2(\overline{x}) = \frac{\sigma^2}{n}$. Si $\frac{n}{N} < 0.10$, le facteur de correction peut également être considéré égal à 1.

Pour les variables binaires, la variabilité d'échantillonnage se définit de la même façon:

$$\sigma^2(p) = \frac{\pi \times (1-\pi)}{n} \frac{N-n}{N-1}$$

Interprétation de la variabilité d'échantillonnage

Plus la variabilité d'échantillonnage est petite, plus l'estimation de la moyenne de la population est précise. Plus la variabilité d'échantillonnage est grande, moins l'estimation de la moyenne de la population est précise.

Remarque: En pratique, comme σ^2 et π ne sont pas connus, on les remplace dans les formules ci-dessus par s^2 et p obtenus à partir de l'échantillon.

Intervalle de confiance à 95%

L'intervalle de confiance à 95% de la moyenne μ de population est une fourchette de valeurs qui recouvre, avec une probabilité de 95%, la vraie valeur de la moyenne μ. On le détermine de la façon suivante:

$$\overline{x} - 1.96 \times \sqrt{\frac{s^2}{n}\frac{N-n}{N-1}} \leq \mu \leq \overline{x} + 1.96 \times \sqrt{\frac{s^2}{n}\frac{N-n}{N-1}}$$

et pour une variable binaire:

$$p - 1.96 \times \sqrt{\frac{p \times (1-p)}{n}\frac{N-n}{N-1}} \leq \pi \leq p + 1.96 \times \sqrt{\frac{p \times (1-p)}{n}\frac{N-n}{N-1}}$$

La précision statistique ou erreur-type (*standard error*) de $\overline{x}$ ou de p, notée $SE(\overline{x})$ ou $SE(p)$, vaut $\sqrt{\frac{s^2}{n}\frac{N-n}{N-1}}$ ou $\sqrt{\frac{p\times(1-p)}{n}\frac{N-n}{N-1}}$ pour une proportion.

Exemples

1. Dans une population de 1000 individus, on a extrait un échantillon d'effectif 200. On a mesuré la pression artérielle systolique (PAS, mmHg) et on a obtenu $\overline{x} = 165$ mmHg et $s = 15$ mmHg. L'intervalle de confiance à 95% pour la moyenne μ de la PAS dans la population est :

$$165 - 1.96 \times \sqrt{\frac{15^2}{200}\frac{800}{999}} \leq \mu \leq 165 + 1.96 \times \sqrt{\frac{15^2}{200}\frac{800}{999}}$$

$$163.1 \leq \mu \leq 166.7$$

On peut donc affirmer avec une confiance de 95% que l'intervalle $[163.1 - 166.7]$ mmHg contient la vraie moyenne μ de la PAS dans la population. L'erreur type de $\overline{x}$ vaut quant à elle 0.95.

2. Dans une population de 10000 individus, on a demandé à 400 personnes si elles fumaient ou non. La proportion de fumeurs était de 35%.
L'intervalle de confiance à 95% pour la moyenne π de la population est :

$$0.35 - 1.96 \times \sqrt{\frac{0.35 \times 0.65}{400}\frac{9600}{9999}} \leq \pi \leq 0.35 + 1.96 \times \sqrt{\frac{0.35 \times 0.65}{400}\frac{9600}{9999}}$$
$$0.304 \leq \pi \leq 0.396$$

On peut donc affirmer avec une confiance de 95% que l'intervalle $[30.4 - 39.6]\%$ contient la vraie proportion π de fumeurs dans la population. L'erreur type de p vaut quant à elle 0.0234 ou 2.3%.

Remarque: Puisque $\frac{400}{10000} < 0.10$, on peut omettre le facteur de correction et utiliser la fomule simplifiée. On a

$$0.35 - 1.96 \times \sqrt{\frac{0.35 \times 0.65}{400}} \leq \pi \leq 0.35 + 1.96 \times \sqrt{\frac{0.35 \times 0.65}{400}}$$
$$0.303 \leq \pi \leq 0.397$$

Les résultats sont fort semblables et l'erreur type de p vaut $\sqrt{\frac{0.35 \times 0.65}{400}} = 0.0238$.

Chapitre 2

Les méthodes d'échantillonnage

Soit une population d'effectif N. Il y a plusieurs façons de tirer des échantillons d'effectif n de cette population. On souhaite tirer un échantillon d'effectif n suffisant pour donner des estimations précises. Dans ce chapitre, on décrit 4 méthodes d'échantillonnage (*sampling methods*): l'échantillonnage simplement fortuit, l'échantillonnage stratifié, l'échantillonnage systématique et l'échantillonnage "en grappes".

2.1 Echantillonnage simplement fortuit

Un échantillonnage est simplement fortuit (*simple random sampling*) s'il répond aux 4 conditions suivantes:

- n doit être fixé à l'avance
- les tirages successifs se font au hasard: tout le monde a la même chance d'être tiré
- et de façon indépendante: un tirage ne conditionne pas un autre tirage
- les tirages successifs se font d'une population invariante : si l'effectif de la population est fini, il faut remettre le sujet tiré dans la population.

Soit une population d'effectif N et X une variable de moyenne μ et de variance σ^2 dans cette population. Si l'on considère un échantillonnage fortuit sans remplacement d'effectif n de cette population, la moyenne de tous les échantillons possibles

d'effectif n extraits de la population vaut μ, ce qu'on écrit $E(\overline{x}) = \mu$, et la variabilité de $\overline{x}$ vaut:

$$Var(\overline{x}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$$

Si X est une variable binaire de moyenne π dans la population, la moyenne de tous les échantillons possibles d'effectif n extraits de la population vaut $E(p) = \pi$ et la variabilité de p vaut:

$$Var(p) = \frac{\pi \times (1-\pi)}{n}\frac{N-n}{N-1}$$

Remarque: Pour extraire des sujets au hasard, il y a plusieurs méthodes: tables de nombres aléatoires (voir Annexe A), générateur de nombres aléatoires (PC), ...

2.2 Echantillonnage stratifié

Une *strate* est une sous-classe de la population. Les populations sont souvent divisées en strates: sexe, race, classes d'âge, classes sociales, classes professionnelles,...

Soit une population d'effectif N comprenant k strates, notées $S_1, \cdots S_k$. Notons N_i le nombre de sujets dans la strate S_i $(i = 1, \cdots, k)$. La somme des N_i est égale à N. Notons enfin μ_i et σ_i^2, la moyenne et la variance de X dans la strate S_i $(i = 1, \cdots, k)$.

Un échantillonnage stratifié (*stratified sampling*) d'effectif n consiste à tirer un échantillon simplement fortuit d'effectif n_i de chaque strate S_i de telle sorte que la somme des n_i égale n. Désignons par $\overline{x}_i$ et s_i^2 la moyenne et la variance de X dans l'échantillon d'effectif n_i $(i = 1, \cdots, k)$.

Soit X la variable étudiée. La moyenne μ de X dans la population peut s'exprimer en fonction des moyennes de X dans les strates.
En effet,

$$\mu = \frac{N_1}{N}\mu_1 + \cdots + \frac{N_k}{N}\mu_k$$

Si on note $\overline{x}_{st}$ la moyenne de l'échantillon stratifié d'effectif n, on a évidemment :

$$\overline{x}_{st} = \frac{N_1}{N}\overline{x}_1 + \cdots + \frac{N_k}{N}\overline{x}_k$$

La moyenne de tous les échantillons stratifiés possibles d'effectif n extraits de la population vaut μ, c'est-à-dire $E(\overline{x}_{st}) = \mu$.

La variabilité d'échantillonnage de $\overline{x}_{st}$ vaut, quant à elle:

$$Var(\overline{x}_{st}) = \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \frac{\sigma_i^2}{n_i} \frac{N_i - n_i}{N_i - 1}$$

Si X est une variable binaire, $E(p_{st}) = \pi$ et

$$Var(p_{st}) = \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \frac{\pi_i(1-\pi_i)}{n_i} \frac{N_i - n_i}{N_i - 1}$$

Choix des n_i

On choisit en général les n_i proportionnels aux N_i $(i = 1, \cdots, k)$. Par contre, s'il s'agit de proportions et que les π_i sont connus approximativement, alors l'échantillonnage optimum est obtenu en choisissant les n_i proportionnels à

$$\frac{N_i\sqrt{\pi_i\,(1-\pi_i)}}{\sum_{i=1}^{k} N_i\sqrt{\pi_i\,(1-\pi_i)}}$$

Exemple

Soit une population composée de 4 hommes et de 4 femmes. Parmi les femmes, une seule fume alors que parmi les hommes, il y a trois fumeurs. On tire un échantillon d'effectif 2 de cette population. Quel type d'échantillonnage est le plus précis, un échantillonnage simplement fortuit ou un échantillonnage stratifié ?

On connaît les données de population:

Femmes	Hommes	Total
$N_1 = 4$	$N_2 = 4$	$N = 8$
$\pi_1 = 0.25$	$\pi_2 = 0.75$	$\pi = 0.50$

Échantillonnage simplement fortuit

Le nombre total d'échantillons simplement fortuits d'effectif $n = 2$ que l'on peut extraire de la population est égal à $C_8^2 = \frac{8!}{2!6!} = 28$.

Il y a 6 échantillons $0-0$ $(p = 0)$, 6 échantillons $1-1$ $(p = 1)$ et 16 échantillons $0-1$ $(p = 0.5)$.

$$E(p) = (6 \times 0 + 6 \times 1 + 16 \times 0.5)/28 = 0.5 = \pi$$

$$Var(p) = \frac{0.5 \times 0.5}{2}\left(\frac{6}{7}\right) = \frac{3}{28} = 0.1071$$

Échantillonnage stratifié
$n_1 = 1$ et $n_2 = 1$.
Il y a 3 échantillons $0-0$ ($p = 0$), 3 échantillons $1-1$ ($p = 1$) et 10 échantillons $0-1$ ($p = 0.5$). Le nombre total d'échantillons stratifiés d'effectif $n = 2$ que l'on peut extraire de la population vaut donc 16.

$$E(p_{st}) = (3 \times 0 + 3 \times 1 + 10 \times 0.5)/16 = 0.5 = \pi$$

$$Var(p_{st}) = \left(\frac{4}{8}\right)^2 \frac{0.25 \times 0.75}{1}\left(\frac{3}{3}\right) + \left(\frac{4}{8}\right)^2 \frac{0.75 \times 0.25}{1}\left(\frac{3}{3}\right) = 0.0938$$

On constate que pour estimer π il vaut mieux avoir recours à un échantillonnage stratifié plutôt qu'à un échantillonnage simplement fortuit puisque la variabilité d'échantillonnage est plus petite ($0.0938 < 0.1071$). Notons au passage que le nombre d'échantillons possibles dans l'échantillonnage stratifié est toujours plus petit que dans le cas d'un échantillonnage simplement fortuit !

2.3 Echantillonnage systématique

Soit une population d'effectif N de laquelle on veut extraire un échantillon d'effectif n. L'échantillonnage systématique (*systematic sampling*) se fait selon le procédé suivant:

- Calculer $r = N/n$. Comme r n'est pas toujours un entier, prendre l'entier r^* le plus proche de r.
- Choisir au hasard un nombre entier, i, compris entre 1 et r^*
- Les sujets de l'échantillon sont les suivants: $i, i + r^*, i + 2r^*, i + 3r^*, \cdots$

Notons que le nombre d'échantillons systématiques possibles est fini et vaut r^*. L'effectif de l'échantillon obtenu n'est pas nécessairement égal à n, cela dépend de la valeur retenue r^*.

Exemple
$N = 8059$ et $n = 300$
$r = 8059/300 = 26.9$
si $r^* = 27$ et $i = 10$, les sujets de l'échantillon sont: $10, 37, 64, 91, \cdots$ $n' = 298$
si $r^* = 26$ et $i = 10$, les sujets de l'échantillon sont: $10, 36, 62, 88, \cdots$ n" $= 309$

Il y a d'autres approches d'échantillonnages systématiques comme par exemple tous les individus nés un 28 mai, ou les maisons ayant le même numéro ou la même position dans une rue.

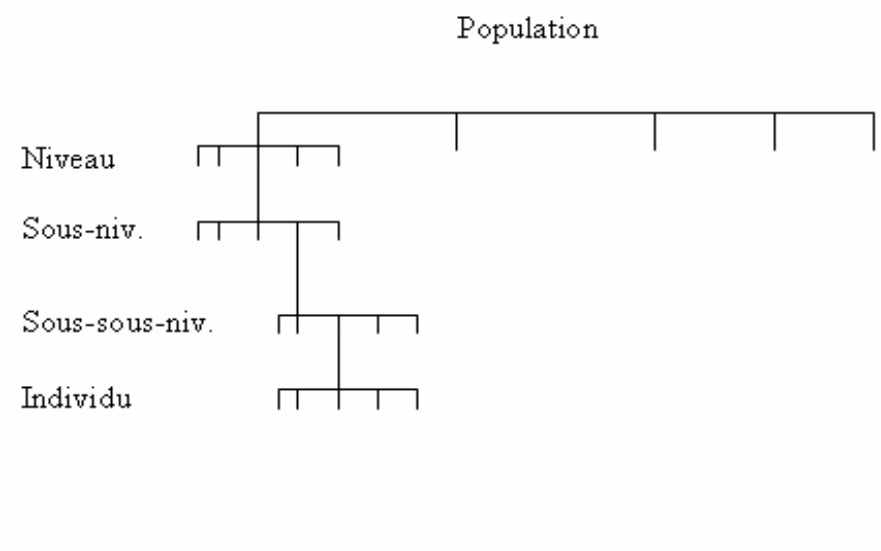
L'échantillonnage systématique a pour avantage d'être simple, rapide et efficace mais surtout de couvrir complètement le domaine échantillonné. Il peut engendrer un biais si les tirages sont liés au problème investigué. Par exemple, si l'on étudie les revenus de personnes et que les habitations échantillonnées sont situées en coin de rue, il y a de fortes chances que ces personnes soient plus aisées, entraînant un biais de sélection.

Si p_{sys} désigne la proportion observée dans un échantillon systématique, alors la moyenne sur tous les échantillons systématiques possibles d'effectif n extraits de la population vaut $E(p_{sys}) = \pi$. Pour la variabilité d'échantillonnage de p_{sys}, on utilise la formule classique $Var(p_{sys}) = \frac{\pi(1-\pi)}{n}\frac{N-n}{N-1}$.

2.4 Echantillonnage en grappe

L'échantillonnage en grappe (*cluster sampling*) est une technique très utilisée. Elle a notamment servi pour l'enquête de santé en Belgique (Health Interview Survey). La population est décomposée en niveaux, et on échantillonne la population à chaque niveau.

Dans l'étude du HIS, l'échantillon était d'effectif n=10,000. Le premier niveau était constitué des municipalités et l'échantillonnage se faisait proportionnellement à la population de ces municipalités (PSU = Primary Sampling Unit). Le second niveau correspondait à un échantillon de ménages par municipalité (SSU = Secondary Sampling Unit). Le troisième niveau était un échantillon de une à quatre personnes par ménage (TSU = Tertiary Sampling Unit).



L'échantillonnage en grappe est pratique et efficace.

Si les grappes diffèrent peu entre elles et qu'il y a hétérogénéité à l'intérieur de celles-ci, la variabilité d'un échantillonnage en grappe est plus petite que la variabilité d'un échantillonnage simplement fortuit.

Si les grappes diffèrent fort entre elles et qu'il y a homogénéité à l'intérieur de celles-ci, la variabilité d'un échantillonnage en grappe est plus grande que la variabilité d'un échantillonnage simplement fortuit.

Chapitre 3

Calcul de la taille d'un échantillon

Le calcul de la taille d'un échantillon est un exercice fréquent et essentiel. L'effectif ne doit pas être trop grand car les études peuvent parfois coûter cher. Il doit cependant être suffisamment grand pour que les tests réalisés sur l'échantillon soient puissants. Le calcul de la taille de l'échantillon s'appelle également calcul de puissance (*power computation*).

3.1 Proportion

3.1.1 N infini

Soit π la proportion à estimer. Combien de sujets (n) doit-on tirer de la population pour que π soit estimé avec une précision donnée au niveau d'incertitude α?

On sait que la moyenne de tous les échantillons possibles d'effectif n extraits de la population vaut π. La variabilité de p vaut $\frac{\pi(1-\pi)}{n}$ dans une population infinie. On peut démontrer que la proportion observée se distribue selon une loi normale de moyenne π et de variance $\frac{\pi(1-\pi)}{n}$. Autrement dit, $p \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$.

On peut dire, qu'avec une probabilité de $(1-\alpha) \times 100\%$, p est compris dans l'intervalle:

$$\left[\pi - Q_G(1-\alpha/2)\sqrt{\frac{\pi(1-\pi)}{n}}, \pi + Q_G(1+\alpha/2)\sqrt{\frac{\pi(1-\pi)}{n}}\right]$$

où $Q_G(1-\alpha/2)$ correspond au quantile $1-\alpha/2$ de la distribution normale (voir Annexe B).

Posons $\Delta = Q_G(1-\alpha/2)\sqrt{\frac{\pi(1-\pi)}{n}}$ la précision souhaitée sur π. En fixant Δ, on peut donc trouver la valeur de n, pour autant que π soit connu approximativement.

On trouve pour n la valeur suivante:

$$n = \left[Q_G(1-\alpha/2)\right]^2 \frac{\pi(1-\pi)}{\Delta^2}$$

Exemple
Supposons que $\pi = 0.30$, $\alpha = 0.05$ et $\Delta = 0.04$, que vaut n ?

$$n = 1.96^2 \frac{0.30 \times (1-0.30)}{0.04^2} = 504$$

Si on tire 504 individus de la population, on est sûr à 95% que la fourchette $[0.26 - 0.34]$ contient la proportion inconnue π ou, en d'autres termes, que la proportion observée ne s'éloigne pas de plus de 4% de π.

3.1.2 N fini

Si on reprend le même raisonnement que pour N infini, en posant $\Delta = Q_G(1-\alpha/2)\sqrt{\frac{\pi(1-\pi)}{n}\frac{N-n}{N-1}}$, on obtient la formule suivante pour n:

$$n = \frac{\left[Q_G(1-\alpha/2)\right]^2 \pi(1-\pi)N}{(N-1)\Delta^2 + \left[Q_G(1-\alpha/2)\right]^2 \pi(1-\pi)}$$

Exemple
Supposons que $N = 1000$, $\pi = 0.30$, $\alpha = 0.05$ et $\Delta = 0.04$, que vaut n ?

$$n = 1.96^2 \frac{0.30 \times (1-0.30) \times 1000}{(1000-1)0.04^2 + 1.96^2 \times 0.30 \times (1-0.30)} = 335$$

Si on tire 335 individus de la population, on est sûr à 95% que la fourchette $[0.26 - 0.34]$ contient la proportion inconnue π ou, en d'autres termes, que la proportion observée ne s'éloigne pas de plus de 4% de π.

3.2 Moyenne

Soit X une variable quantitative continue. Supposons qu'elle est distribuée suivant une loi normale de moyenne μ et de variance σ^2. Que vaut l'effectif n de l'échantillon pour que la moyenne d'échantillon $\overline{x}$ ne s'écarte pas plus de Δ de la moyenne réelle μ au niveau d'incertitude α ? On procède exactement de la même façon que pour déterminer l'effectif de l'échantillon pour une proportion.

3.2.1 N infini

$$n = \frac{\left[Q_G(1-\alpha/2)\right]^2 \sigma^2}{\Delta^2}$$

3.2.2 N fini

$$n = \frac{\left[Q_G(1-\alpha/2)\right]^2 \sigma^2 N}{(N-1)\Delta^2 + \left[Q_G(1-\alpha/2)\right]^2 \sigma^2}$$

Exemple
Supposons que $N = 1000$,$\sigma^2 = 3.5$,$\alpha = 0.05$ et $\Delta = 0.8$, que vaut n ?

$$n = \frac{1.96^2 \times 3.5 \times 1000}{(1000-1) \times 0.8^2 + 1.96^2 \times 3.5} = 21$$

Si on tire 21 individus de la population, on est sûr à 95% que la moyenne observée ne s'éloignera pas de plus de 0.8 unités de la vraie moyenne μ. Il y a peu d'individus dans l'échantillon car Δ est grand, il représente presque la moitié de $\sigma = 1.871$.

3.3 Comparaison de deux proportions (π_1 et π_2)

Supposons que l'on ait deux populations et que l'on souhaite comparer une proportion entre ces populations.
Soient π_1 et π_2 et n_1 et n_2 les proportions et les effectifs d'échantillons pour chaque population.

L'hypothèse à tester est:

H_0:$\pi_1 = \pi_2$
H_1:$\pi_1 \neq \pi_2$

ou encore, si l'on pose $\Delta = \pi_1 - \pi_2$
H_0:$\Delta = 0$
H_1:$\Delta \neq 0$

Le problème consiste à trouver n_1 et n_2 tels que si H_0 est vraie, la probabilité de rejeter H_0 est au plus $\alpha = 0.05$ et si H_0 est fausse, la probabilité d'accepter H_0 est au plus $\beta = 0.10$ ou 0.20.

Supposons $n_1 = n_2 = n$

L'effectif n est obtenu en résolvant l'équation suivante:

$$\Delta = Q_G(1-\alpha/2)\sqrt{\frac{2\pi_1(1-\pi_1)}{n}} + Q_G(1-\beta)\sqrt{\frac{\pi_1(1-\pi_1)}{n} + \frac{\pi_2(1-\pi_2)}{n}}$$

Exemple
Supposons qu'en France, l'incidence d'une maladie soit de 3% ($\pi_1 = 0.03$) et qu'en Belgique, l'incidence de cette maladie soit de 2% ($\pi_2 = 0.02$). De quelle taille doit être l'échantillon en Belgique et en France si on veut détecter une différence Δ de 1% (0.01) avec une probabilité de 95% ($\alpha = 0.05$) et une puissance de 90% ($\beta = 0.10$) ?

Il faut résoudre l'équation suivante:

$$\Delta = 1.96\sqrt{\frac{2 \times 0.03 \times 0.97}{n}} + 1.28\sqrt{\frac{0.03 \times 0.97}{n} + \frac{0.02 \times 0.98}{n}}$$

On trouve alors $n = n_1 = n_2 = 5705$.

Remarque: Parfois, on souhaite que $n_1 = kn$ ($k \geq 1$) et $n_2 = n$. C'est le cas des études cas /contrôle. La formule précédente devient alors:

$$\Delta = Q_G(1-\alpha/2)\sqrt{\frac{\pi_1(1-\pi_1)}{kn} + \frac{\pi_1(1-\pi_1)}{n}} + Q_G(1-\beta)\sqrt{\frac{\pi_1(1-\pi_1)}{kn} + \frac{\pi_2(1-\pi_2)}{n}}$$

3.4 Comparaison de deux moyennes (μ_1 et μ_2)

Soit X une variable quantitative telle que
$X \sim N(\mu_1, \sigma_1^2)$ dans la population P_1
$X \sim N(\mu_2, \sigma_2^2)$ dans la population P_2

Soit $\Delta = \mu_1 - \mu_2$
L'effectif n de chacun des échantillons s'obtient à partir de la formule suivante:

$$\Delta = Q_G(1-\alpha/2)\sqrt{\frac{2\sigma_1^2}{n}} + Q_G(1-\beta)\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}}$$

Dans le cas où $\sigma_1^2 = \sigma_2^2 = \sigma^2$, la formule devient:

$$n = \frac{2\sigma^2 \left[Q_G(1-\alpha/2) + Q_G(1-\beta)\right]^2}{\Delta^2}$$

Exemple

Pour $\alpha = 0.05$ et $\beta = 0.10$, si $\Delta = 1$ et $\sigma = 1.8$, $n = n_1 = n_2 = 68$.

Chapitre 4

Risque relatif et odds ratio

Le problème est de mesurer le degré d'association entre une maladie (survenance ou existence) et un (ou plusieurs) facteurs de risque. La méthode utilisée est celle du risque relatif ou encore de l'odds ratio.

Notations

M: maladie	M_+ : présente
	M_- : absente
F: facteur de risque	F_+ : exposé
	F_- : non-exposé

4.1 Risque relatif (RR)

4.1.1 Etudes prospectives

Considérons une population constituée de sujets exposés (F_+) et de sujets non-exposés (F_-) à un risque donné. On suit ces sujets *prospectivement* dans le temps et on comptabilise le nombre de sujets qui développent la maladie (M_+) et ceux qui ne la développent pas (M_-). Au départ, on a exclu les sujets qui avaient déjà la maladie, autrement dit, on a exclu les cas de *prévalence*. Les sujets qui développent la maladie M sont appelés les cas d'*incidence*.

En théorie, la population est répartie de la façon suivante:

		Maladie		
		M_+	M_-	
Facteur de risque	F_+	A	B	$A+B$
	F_-	C	D	$C+D$
		$A+C$	$B+D$	N

fixé

observé

Dans la population fixée au départ, on ne connaît pas le nombre de sujets qui vont développer la maladie, c'est pourquoi $A+C$ et $B+D$ sont observés.

Définition du Risque Relatif (*RR*)
$\frac{A}{A+B}$ est le taux d'incidence de la maladie M dans le groupe exposé au facteur de risque.
$\frac{C}{C+D}$ est le taux d'incidence de la maladie M dans le groupe non-exposé au facteur de risque.

Le *risque relatif* (*relative risk*) est le rapport entre le taux d'incidence de la maladie M dans le groupe exposé et le taux d'incidence de la maladie M dans le groupe non-exposé au facteur de risque. Autrement dit, le risque relatif est la proportion de sujets qui développent la maladie M dans le groupe exposé sur celle dans le groupe non-exposé. C'est un nombre positif (ou nul).

$$RR = \frac{\frac{A}{A+B}}{\frac{C}{C+D}}$$

Si $RR > 1$, on dit qu'il y a une association positive entre le facteur de risque et la maladie.
Si $RR < 1$, il y a une association négative entre le facteur de risque et la maladie. Le fait d'être exposé au facteur de risque protège de la maladie.
Si $RR = 1$, il n'y a pas d'association entre le facteur de risque et la maladie.

En pratique, on mesure l'association entre le facteur de risque et la maladie sur un échantillon d'effectif n extrait de la population. On obtient donc un tableau de la forme suivante:

Facteur de risque		Maladie M_+	Maladie M_-	
	F_+	a	b	$a+b$
	F_-	c	d	$c+d$
		$a+c$	$b+d$	n

fixé (colonnes) — **observé** (lignes)

Le risque relatif RR est estimé par

$$\hat{RR} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

Exemple
Honolulu Heart Programme (Abbot, 1980)
Maladie: thrombose cérébrale
Facteur de risque: tabac
Etude prospective qui a duré 12 ans.

Tabac		Thrombose cérébrale M_+	Thrombose cérébrale M_-	
	F_+	171	3264	3435
	F_-	117	4320	4437
		288	7584	7872

Le taux d'incidence de thrombose cérébrale chez les fumeurs vaut : $\frac{171}{3435} = 0.0498$.
Le taux d'incidence de thrombose cérébrale chez les non-fumeurs vaut : $\frac{117}{4437} = 0.0264$.

$$\hat{RR} = \frac{0.0498}{0.0264} = 1.89$$

Le risque relatif obtenu est supérieur à 1, cela signifie que le fait de fumer double quasiment le risque de développer une thrombose cérébrale.

4.1.2 Etudes transversales

Considérons la population à un moment donné. Celle-ci est constituée de sujets exposés (F_+) et de sujets non-exposés (F_-) à un facteur de risque. A ce même moment, certains sujets ont la maladie M alors que d'autres ne l'ont pas. Une étude *transversale* (*cross-sectional study*) consiste à observer, à un moment donné, une

population et de comptabiliser les sujets qui sont exposés à un facteur de risque et ceux qui ne le sont pas et simultanément ceux qui ont la maladie M et ceux qui ne l'ont pas.

En théorie, la population est répartie de la façon suivante:

		Maladie		
		M_+	M_-	
Facteur de risque	F_+	A	B	$A+B$
	F_-	C	D	$C+D$
		$A+C$	$B+D$	N

observé (ligne) — **observé** (colonne)

Dans la population que l'on observe, on ne connaît pas à l'avance le nombre de sujets qui ont la maladie ni ceux qui sont exposés.

Définition du Risque Relatif (*RR*)
$\frac{A}{A+B}$ est le taux de prévalence de la maladie M dans le groupe exposé au facteur de risque.
$\frac{C}{C+D}$ est le taux de prévalence de la maladie M dans le groupe non-exposé au facteur de risque.

Le risque relatif est le rapport des prévalences

$$RR = \frac{\frac{A}{A+B}}{\frac{C}{C+D}}$$

En pratique, on mesure l'association entre le facteur de risque et la maladie sur un échantillon d'effectif n extrait de la population.
Le risque relatif RR est estimé par $\hat{RR}$

$$\hat{RR} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

4.1.3 Etudes rétrospectives

Considérons une population constituée de sujets atteints de la maladie (M_+) et de sujets non-atteints (M_-). On observe ces sujets rétrospectivement. En clair, on va

comptabilier ceux qui étaient exposés et ceux qui n'étaient pas exposés à un facteur de risque F.

En théorie, dans une étude rétrospective (*case-control study*), la population est répartie de la façon suivante:

		Maladie		
		M_+	M_-	
Facteur de risque	F_+	A	B	$A+B$
	F_-	C	D	$C+D$
		$A+C$	$B+D$	N

observé

fixé

Dans la population fixée au départ, on ne connaît pas le nombre de sujets qui ont été exposés ou non à un facteur de risque, c'est pourquoi $A+B$ et $C+D$ sont observés.

En pratique, cela n'a plus de sens de calculer $\frac{a}{a+b}$ ni $\frac{c}{c+d}$ car $\frac{a}{a+b}$ (respectivement $\frac{c}{c+d}$) n'estime plus $\frac{A}{A+B}$ (respectivement $\frac{C}{C+D}$). **Dans les études rétrospectives, cela n'a aucun sens de parler de risque relatif !**

Remarque

Il y a cependant deux cas où on peut utiliser le risque relatif dans les études rétrospectives. Dans ces deux cas, le risque relatif se réduit à un odds ratio que l'on peut utiliser dans les études rétrospectives.

1. Lorsque la maladie M est très rare

   Dans ce cas, $A+C$ est très petit, il en va donc de même pour A et C. Par conséquent, B et D sont très grands. On peut donc *assimiler* $A+B$ à B et $C+D$ à D.

   $$RR \approx \frac{A/B}{C/D} = \frac{AD}{BC} = \frac{A/C}{B/D}$$

   On peut alors estimer le risque relatif puisque à présent a/c estime bien A/C et b/d estime bien B/D.

2. Lorsque la population M_- représente toute la population

   Dans ce cas, $A + B = B$ et $C + D = D$. Dès lors, $RR = \frac{A/B}{C/D} = \frac{A/C}{B/D}$ et par le même raisonnement que ci-dessus, le risque relatif est estimé par $\frac{a/c}{b/d}$.

4.2 Odds ratio (*OR*)

L'*odds ratio* s'utilise indifféremment dans les études prospectives, transversales et rétrospectives. On appelle également l'odds ratio, le *rapport croisé*.

Un odds est une sorte de pari. Par exemple, dans les courses hippiques, on dit qu'un cheval est coté à 4 contre 1. Cela signifie que la probabilité qu'il a de gagner est de $4/(4 + 1)$.

En théorie, la population est répartie de la façon suivante:

		Maladie		
		M_+	M_-	
Facteur de risque	F_+	A	B	$A + B$
	F_-	C	D	$C + D$
		$A + C$	$B + D$	N

Définition de l'Odds Ratio (*OR*)
$\frac{A}{B}$ est l'odds sur la maladie dans le groupe exposé
$\frac{C}{D}$ est l'odds sur la maladie dans le groupe non-exposé

L'odds ratio est le rapport des odds du groupe exposé et du groupe non-exposé.

$$OR = \frac{A/B}{C/D} = \frac{AD}{BC} = \frac{A/C}{B/D}$$

C'est une mesure d'association entre la maladie M et le facteur de risque F. C'est un nombre toujours positif (ou nul).
Si $OR > 1$, l'association entre la maladie et le facteur de risque est positive.

Si $OR < 1$, l'association entre la maladie et le facteur de risque est négative.
Si $OR = 1$, il n'y a pas d'association entre la maladie et le facteur de risque.

En pratique, l'odds ratio est estimé par $\hat{OR} = \frac{ad}{bc}$.

Remarque
L'estimation de l'odds ratio est toujours plus forte que l'estimation du risque relatif.
Si $\hat{RR} > 1$, alors $\hat{OR} > \hat{RR} > 1$
Si $\hat{RR} < 1$, alors $\hat{OR} < \hat{RR} < 1$

Si les lignes (resp. les colonnes) de la table sont permutées, l'estimation de l'odds ratio est inversée ($1/OR$). Si les lignes et les colonnes de la table sont permutées, l'odds ratio reste le même.

Exemple
Honolulu Heart Programme (Abbot, 1980)

$$\hat{RR} = \frac{0.0498}{0.0264} = 1.89$$
$$\hat{OR} = \frac{171 \times 430}{117 \times 3264} = 1.93 > \hat{RR} > 1$$

4.3 Intervalles de confiance pour RR et OR

L'utilisation d'intervalles de confiance pour le risque relatif ou pour l'odds ratio permet:

- d'estimer la précision statistique des estimations de RR et de OR
- de tester l'hypothèse nulle qu'il n'y a pas d'association entre la maladie et le facteur de risque (H_0: $OR = 1$ ou $RR = 1$).

4.3.1 Intervalle de confiance pour OR

H_0: $OR = 1$
H_1: $OR \neq 1$

Tester cette hypothèse peut se faire l'aide des intervalles de confiance. Si l'intervalle de confiance ne contient pas la valeur 1, on rejette l'hypothèse nulle. Dans le cas contraire, on ne rejette pas l'hypothèse nulle.

L'intervalle de confiance pour un OR se calcule par la *méthode de Woolf*.

1. On calcule $\hat{OR} = \dfrac{ad}{bc}$
2. On calcule $ln\left(\hat{OR}\right)$
3. On calcule la variance de $ln\left(\hat{OR}\right)$

   $$Var\left[ln\left(\hat{OR}\right)\right] = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

   L'erreur type de $ln\left(\hat{OR}\right)$ correspond à la racine carrée de la variance et est noté $SE\left[ln\left(\hat{OR}\right)\right] = \sqrt{Var\left[ln\left(\hat{OR}\right)\right]}$
4. On calcule un intervalle de confiance à 95% pour $ln(OR)$

   $$ln\left(\hat{OR}_1\right) = ln\left(\hat{OR}\right) - 1.96 \times SE\left[ln\left(\hat{OR}\right)\right]$$

   $$ln\left(\hat{OR}_2\right) = ln\left(\hat{OR}\right) + 1.96 \times SE\left[ln\left(\hat{OR}\right)\right]$$
5. L'intervalle de confiance à 95% pour OR est donné par $\hat{OR}_1 = e^{ln\left(\hat{OR}_1\right)}$ et $\hat{OR}_2 = e^{ln\left(\hat{OR}_2\right)}$, d'où $\hat{OR}_1 \leq OR \leq \hat{OR}_2$.

Exemple
Y a-t-il une association significative entre le diabète (facteur de risque) et la cataracte (maladie) ?

		Cataracte		
		M_+	M_-	
Diabète	F_+	55	84	
	F_-	552	1927	
		607	2011	2618

1. $\hat{OR} = \dfrac{55 \times 1927}{84 \times 552} = 2.29$
2. $ln\left(\hat{OR}\right) = 0.8267$
3. $Var\left[ln\left(\hat{OR}\right)\right] = \frac{1}{55} + \frac{1}{84} + \frac{1}{552} + \frac{1}{1927} = 0.0324$
   $SE\left[ln\left(\hat{OR}\right)\right] = \sqrt{0.0324} = 0.18$
4. $ln\left(\hat{OR}_1\right) = 0.8267 - 1.96 \times 0.18 = 0.4738$
   $ln\left(\hat{OR}_2\right) = 0.8267 + 1.96 \times 0.18 = 1.1796$
5. L'intervalle de confiance à 95% pour OR est donné par $e^{0.4738}$ et $e^{1.1786}$ et on a $1.6 \leq OR \leq 3.3$.

Cet intervalle ne contient pas la valeur 1. Il y a donc une association significative entre le diabète et la cataracte. Les personnes qui ont le diabète ont un risque deux fois plus élevé d'avoir la cataracte.

4.3.2 Intervalle de confiance pour RR

H_0: $RR = 1$
H_1: $RR \neq 1$

Tester cette hypothèse peut se faire l'aide des intervalles de confiance. Si l'intervalle de confiance ne contient pas la valeur 1, on rejette l'hypothèse nulle. Dans le cas contraire, on ne rejette pas l'hypothèse nulle.

L'intervalle de confiance pour un RR se calcule par la *méthode de Katz*.

1. On calcule $\hat{RR} = \dfrac{\frac{a}{a+b}}{\frac{c}{c+d}}$
2. On calcule $ln\left(\hat{RR}\right)$
3. On calcule la variance de $ln\left(\hat{RR}\right)$

$$Var\left[ln\left(\hat{RR}\right)\right] = \frac{b/a}{a+b} + \frac{d/c}{c+d}$$

   L'erreur type de $ln\left(\hat{RR}\right)$ correspond à la racine carrée de la variance et est noté $SE\left[ln\left(\hat{RR}\right)\right] = \sqrt{Var\left[ln\left(\hat{RR}\right)\right]}$

4. On calcule un intervalle de confiance à 95% pour $ln(RR)$

$$ln\left(\hat{RR}_1\right) = ln\left(\hat{RR}\right) - 1.96 \times SE\left[ln\left(\hat{RR}\right)\right]$$

$$ln\left(\hat{RR}_2\right) = ln\left(\hat{RR}\right) + 1.96 \times SE\left[ln\left(\hat{RR}\right)\right]$$

5. L'intervalle de confiance à 95% pour RR est donné par $\hat{RR}_1 = e^{ln\left(\hat{RR}_1\right)}$ et $\hat{RR}_2 = e^{ln\left(\hat{RR}_2\right)}$, d'où $\hat{RR}_1 \leq RR \leq \hat{RR}_2$.

Exemple
Y a-t-il une association significative entre le cholestérol (facteur de risque) et l'infarctus myocardique (maladie) ?

		Infarctus myocardique		
		M_+	M_-	
Cholestérol	≥ 250 F_+	10	125	135
	< 250 F_-	21	449	470
		31	574	605

1. $\hat{RR} = \frac{\frac{10}{135}}{\frac{21}{470}} = 1.66$
2. $ln\left(\hat{RR}\right) = 0.5055$
3. $Var\left[ln\left(\hat{RR}\right)\right] = \frac{125/10}{135} + \frac{449/21}{470} = 0.138$
   $SE\left[ln\left(\hat{RR}\right)\right] = 0.3716$
4. $ln\left(\hat{RR}_1\right) = 0.5055 - 1.96 \times 0.3716 = -0.22$
   $ln\left(\hat{RR}_2\right) = 0.5055 + 1.96 \times 0.3716 = 1.23$
5. L'intervalle de confiance à 95% pour RR est donné par $e^{0.22}$ et $e^{1.23}$, d'où on obtient $0.8 \leq RR \leq 3.4$.

Cet intervalle contient la valeur 1. Il n'y a donc pas d'association significative entre le cholestérol et l'infarctus myocardique.

4.4 Risque attribuable

Le but du calcul du *risque attribuable* (*attributable risk*) est de pouvoir identifier les facteurs de risque qui ont réellement un impact sur la survenue d'une maladie M. Le risque attribuable tient compte de la proportion de sujets exposés au facteur de risque F.

Supposons, par exemple, que le risque relatif associé à un facteur de risque F_1 vaut $RR_1 = 8$ et que la proportion de sujets exposés à F_1 vaut π_1. De même, supposons que le risque relatif associé au facteur de risque F_2 vaut $RR_2 = 2$ et que la proportion de sujets exposés à F_2 vaut π_2. Ce n'est pas nécessairement sur le facteur de risque où le risque relatif est le plus élevé qu'il faut agir car la proportion de sujets concernés par ce risque n'est peut-être pas la plus importante.

Considérons une étude prospective. L'échantillon se répartit de la façon suivante:

		Maladie		
		M_+	M_-	
Facteur de risque	F_+	a	b	$a+b$
	F_-	c	d	$c+d$
		$a+c$	$b+d$	n

fixé

observé

$I_1 = \dfrac{a}{a+b}$: taux d'incidence de la maladie dans le groupe exposé

$I_0 = \dfrac{c}{c+d}$: taux d'incidence de la maladie dans le groupe non-exposé

$I_t = \dfrac{a+c}{n}$: taux d'incidence de la maladie dans l'échantillon total

$p =$ estimation de la proportion de sujets exposés dans la population générale.

Notons que $p = \dfrac{a+b}{n}$ si on a affaire à une étude transversale ou rétrospective.

4.4.1 Risque attribuable dans le groupe exposé

Le *risque attribuable dans le groupe exposé* est l'excès de maladie réellement dû au facteur de risque. Il se note AR. Il représente le pourcentage de sujets atteints de la maladie M, parmi les sujets exposés et qui ont la maladie M parce qu'ils étaient exposés.

$$AR = \frac{I_1 - I_0}{I_1}$$

De manière plus générale, on l'exprime en pourcentage

$$AR(\%) = \frac{I_1 - I_0}{I_1} \times 100 = \frac{RR - 1}{RR} \times 100$$

où $RR = \frac{I_1}{I_0}$ est supposé ≥ 1.

4.4.2 Risque attribuable dans la population globale

Le *risque attribuable dans la population globale* est le pourcentage de sujets atteints par la maladie M dans la population générale, qui peut être expliqué par le facteur de risque. Il se note AR_p.

Exprimé en %, le risque attribuable s'écrit :

$$AR_p = \frac{I_t - I_0}{I_t}$$

ou encore

$$AR_p(\%) = \frac{p \times (RR - 1)}{1 + p \times (RR - 1)} \times 100$$

En effet, notons que $I_t = p \times I_1 + (1 - p) \times I_0$. Dès lors, on a successivement :

$$AR_p = \frac{I_t - I_0}{I_t} = \frac{p \times I_1 + (1 - p) \times I_0 - I_0}{p \times I_1 + (1 - p) \times I_0}$$

Exemple

Etude de Kahn et al (1966). Relation entre le tabac et le décès

		Issue		
		Décès M_+	Vie M_-	
Tabac	Oui F_+	1116	700652	701768
	Non F_-	426	1015573	1015999
		1542	1716225	1717767

fixé

observé

$p = \frac{701768}{1717767} = 0.41$ =proportion de sujets exposés (ici fumeurs)

$I_1 = \frac{1116}{701768} = 0.00159$

$I_0 = \frac{426}{1015999} = 0.00042$

$I_t = \frac{1542}{1717767} = 0.000898$

$RR = \frac{I_1}{I_0} = 3.79$

$$AR(\%) = 100 \times \frac{3.79 - 1}{3.79} = 74\%$$

Il y a donc 74% des personnes décédées parmi les fumeurs qui sont décédées parce qu'elles fumaient.

$$AR_p(\%) = 100 \times \frac{0.000898 - 0.00042}{0.000898} = 53\%$$

53% des décès dans la population générale peuvent être expliqués par le tabagisme.

Chapitre 5

Facteurs confondants

Un *facteur confondant* (*confounding factor*) est un facteur susceptible d'altérer l'association entre un facteur de risque et une maladie. L'effet du facteur confondant doit donc être éliminé.

Un facteur est dit confondant si

- il est associé à la maladie
- il est associé au facteur de risque

Exemple
Le risque de mortalité et les cheveux gris. Chacun de ces facteurs est associé à l'âge. En effet, plus l'âge augmente, plus le risque de mortalité augmente. Il en va de même pour les cheveux gris. L'association entre risque de mortalité et cheveux gris est donc *biaisée* par l'âge. Autres facteurs confondants: sexe, race, ...

5.1 Elimination d'un facteur confondant

Notations

M: Maladie M_+: malade
M_-: non-malade

F: Facteur de risque F_+: exposé
F_-: non-exposé

FC: Facteur confondant k = nombre de modalités de FC
FC_1: modalité 1
FC_2: modalité 2

FC_k: modalité k

Pour chaque modalité du facteur confondant, on construit la table $F \times M$. Il y aura donc k tables $F \times M$ $(i = 1, \cdots, k)$.

		Maladie		
		M_+	M_-	
Facteur de risque	F_+	a_i	b_i	
	F_-	c_i	d_i	
				t_i

5.1.1 Méthode de Mantel-Haenszel

Lorsqu'on a vérifié que le facteur est bien un facteur confondant (association avec M et avec F), la méthode de Mantel-Haenszel permet de calculer un odds ratio qui détermine l'association entre F et M en éliminant l'effet du facteur confondant. Cet odds ratio est noté OR_{MH}. Une estimation de cet odds ratio est donnée par:

$$\hat{OR}_{MH} = \frac{\frac{a_1 d_1}{t_1} + \frac{a_2 d_2}{t_2} + \cdots + \frac{a_k d_k}{t_k}}{\frac{b_1 c_1}{t_1} + \frac{b_2 c_2}{t_2} + \cdots + \frac{b_k c_k}{t_k}}$$

ou

$$\hat{OR}_{MH} = \frac{\sum_{i=1}^{k} \frac{a_i d_i}{t_i}}{\sum_{i=1}^{k} \frac{b_i c_i}{t_i}}$$

où a_i, b_i, c_i, d_i et t_i sont les éléments des tables $F \times M$ correspondant aux différentes modalités i du facteur confondant FC.

Exemple Dans une étude rétrospective, on a étudié l'association entre la PAS et l'infarctus du myocarde. Sans tenir compte de l'âge, voici la table obtenue:

		Infactus du myocarde		
		M_+	M_-	
Facteur de risque: PAS	F_+: $PAS > 140$	29	711	740
	F_-: $PAS \leq 140$	27	1244	1271
		56	1955	2011

Si on ne tient pas compte de l'âge, $\hat{OR} = \frac{29 \times 1244}{27 \times 711} = 1.88$.

L'âge est-il un facteur confondant ?

- *Association entre âge et infarctus du myocarde* (maladie)

		Infactus du myocarde		
		M_+	M_-	
Facteur confondant: âge	FC_+: $age > 60$	15	188	
	FC_-: $age \leq 60$	41	1767	
		56	1955	

L'odds ratio vaut 3.44. L'association est significative.

- *Association entre âge et PAS (facteur de risque)*

		PAS		
		F_+	F_-	
Facteur confondant: age	FC_+: $age > 60$	124	79	
	FC_-: $age \leq 60$	616	1192	
		740	1271	

L'odds ratio vaut 3.04. L'association est significative.
L'âge est donc un facteur confondant.

- *OR de Mantel-Haenszel*
Considérons les OR dans les différentes modalités de l'âge ($k = 2$).

	Age > 60		
1	M_+	M_-	
F_+:	$a_1 = 9$	$b_1 = 115$	
F_-:	$c_1 = 6$	$d_1 = 73$	
			$t_1 = 203$

	Age ≤ 60		
2	M_+	M_-	
F_+:	$a_2 = 20$	$b_2 = 596$	
F_-:	$c_2 = 21$	$d_2 = 1171$	
			$t_2 = 1808$

$$\hat{OR}_{MH} = \frac{(9 \times 73/73) + (20 \times 1171/1808)}{(115 \times 6/203) + (21 \times 596/1808)} = 1.57$$

Cette valeur mesure l'association entre la PAS et l'infarctus du myocarde après avoir tenu compte de l'effet de l'âge.

5.1.2 Méthode de Woolf

Lorsqu'on a vérifié que le facteur est bien un facteur confondant (association avec M et avec F), la méthode de Woolf permet de calculer un odds ratio qui détermine l'association entre F et M en éliminant l'effet du facteur confondant. Cet odds ratio est noté OR_W.
Une estimation de cet odds ratio est obtenue de la façon suivante:

1. calculer $ln(\hat{OR}_1)$,$ln(\hat{OR}_2)$, ...,$ln(\hat{OR}_k)$, où les quantités $\hat{OR}_i$ sont les OR des tables $F \times M$ correspondant aux différentes modalités i du facteur confondant FC $(i = 1, \cdots, k)$.

2. calculer pour $i = 1, \cdots, k$:
$$var\left[ln\left(\hat{OR}_i\right)\right] = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$$
où a_i, b_i, c_i, d_i sont les éléments des tables $F \times M$ correspondant aux différentes modalités i du facteur confondant FC.

3. calculer pour $i = 1, \cdots, k$:
$$\omega_i = \frac{1}{var\left[ln\left(\hat{OR}_i\right)\right]}$$

4. $ln(\hat{OR}_W) = \dfrac{\sum_{i=1}^{k} \omega_i ln(\hat{OR}_i)}{\sum_{i=1}^{k} \omega_i}$

5. $\hat{OR}_W = e^{ln(\hat{OR}_W)}$

Remarque
Il est possible de calculer un intervalle de confiance pour l'odds ratio de Woolf. La variance du logarithme de l'estimation de OR_W est donnée par:

$$Var\left[ln(\hat{OR}_W)\right] = \frac{1}{\sum_{i=1}^{k} \omega_i}$$

Les limites de l'intervalle de confiance à 95% pour $ln(OR_W)$ sont calculées de la manière suivante:

$$ln(\hat{OR}_W) \pm 1.96 \times SE\left[ln(\hat{OR}_W)\right]$$

où

$$SE\left[ln\left(\hat{OR}_W\right)\right] = \sqrt{Var\left[ln\left(\hat{OR}_W\right)\right]} = \sqrt{\left(\sum_{i=1}^{k} w_i\right)^{-1}}$$

Il suffit de calculer l'exponentielle de chacune des limites trouvées pour obtenir l'intervalle de confiance à 95% pour OR_W.

Exemple
Dans l'exemple précédent, OR de Woolf:

Age $>$ 60 ans	Age $\leq$ 60 ans
$\hat{OR}_1 = 0.95$	$\hat{OR}_2 = 1.87$
$ln(\hat{OR}_1) = -0.051$	$ln(\hat{OR}_2) = 0.6259$
$Var\left[ln(\hat{OR}_1)\right] = 0.3002$	$Var\left[ln(\hat{OR}_2)\right] = 0.1002$
$\omega_1 = 3.33$	$\omega_2 = 9.98$

$$ln(\hat{OR}_W) = \frac{3.33 \times (-0.051) + 9.98 \times (0.6259)}{3.33 + 9.98} = 0.4576$$

$$\hat{OR}_W = e^{0.4576} = 1.58$$

L'intervalle de confiance à 95% pour OR_W est obtenu en calculant successivement

$Var\left[ln\left(\hat{OR}\right)\right] = \frac{1}{3.33+9.98} = 0.0751$

$SE\left[ln\left(\hat{OR}\right)\right] = \sqrt{0.0751} = 0.274$

$0.4576 \pm 1.96 \times 0.274$

$-0.080 \leq lnOR \leq 0.99$

$e^{-0.080} \leq OR \leq e^{0.99}$

OR_W est donc compris entre 0.92 et 2.70. Il n'y a donc pas d'association entre la PAS et l'infarctus myocardique, après avoir tenu compte de l'âge.

5.2 Test d'interaction

Le test d'interaction consiste à tester si les odds ratios mesurant l'association entre le facteur de risque et la maladie sont les mêmes pour toutes les modalités du facteur confondant.

H_0: $OR_1 = OR_2 = \cdots = OR_k$
H_1: Il existe $i \neq j$ tels que $OR_i \neq OR_j$

Le critère

$$\chi^2_{obs} = \sum_{i=1}^{k} \omega_i \left[ln(\hat{OR}_i) - ln(\hat{OR}_W) \right]^2$$

est distribué sous H_0 comme un chi-carré à $k-1$ degrés de liberté.

Si $\chi^2_{obs} \leq Q_{\chi^2}(0.95; 1)$ alors on accepte H_0. Il n'y a donc pas d'interaction du facteur confondant entre maladie et facteur de risque.
Si $\chi^2_{obs} > Q_{\chi^2}(0.95; 1)$ alors on n'accepte pas H_0. Le facteur confondant interagit sur la maladie et sur le facteur de risque.

Exemple
Dans l'exemple précédent

H_0: $OR_{age>60} = OR_{age\leq 60}$
H_1: $OR_{age>60} \neq OR_{age\leq 60}$

$\chi^2_{obs} = 3.33 \left[ln(0.95) - ln(1.58) \right]^2 + 9.98 \left[ln(1.87) - ln(1.58) \right]^2 = 1.15$

Comme $Q_{\chi^2}(0.95; 1) = 3.84$ (voir Annexe C) et que $\chi^2_{obs} = 1.15 < 3.84$, il n'y a pas d'interaction du facteur confondant entre la PAS et l'infarctus myocardique.

Chapitre 6

Régressions multiple et logistique

6.1 Rappel sur la régression simple

La *régression simple* consiste à exprimer la moyenne d'une variable aléatoire *dépendante* en fonction d'une variable *indépendante*.

Y: variable dépendante (continue)
X: variable indépendante ou explicative

On considère le modèle:

$$E(Y|x) = \beta_0 + \beta_1 x$$

$E(Y|x)$ est l'espérance mathématique de Y conditionnelle à $X = x$. C'est la valeur moyenne de la variable Y pour une valeur donnée x de la variable X. Elle correspond à l'équation de la droite de régression. On appelle β_0 l'ordonnée à l'origine (*intercept*) de la droite de régression et β_1 la pente (*slope*) de la droite de régression.

En pratique, on estime β_0 et β_1 à partir des données. On obtient dès lors une estimation de la moyenne de la variable Y, elle est notée $\hat{Y}$. C'est la valeur prédite de Y à partir de $X = x$.

$$\hat{Y} = b_0 + b_1 x$$

La différence $Y_{obs} - \hat{Y}$ est appelée *résidu* et ne dépend en principe plus de X.

Les valeurs b_0 et b_1 sont calculées sur base du principe des moindres carrés.

6.2 Régression multiple

La *régression multiple* consiste à exprimer la moyenne d'une variable *dépendante* en fonction de plusieurs variables *indépendantes* (explicatives).
Y: variable dépendante (continue)
$X_1, \cdots, X_p$: variables indépendantes ou explicatives

On obtient le modèle:

$$E\left(Y|\underset{\sim}{x}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

$E\left(Y|\underset{\sim}{x}\right)$ est l'espérance mathématique de Y conditionnelle à $\underset{\sim}{X} = \underset{\sim}{x}$. C'est la valeur moyenne de la variable Y pour les valeurs $x_1, \cdots, x_p$. Elle correspond à l'équation de la régression multiple de Y sur $\underset{\sim}{x}$. On appelle β_0 le terme indépendant ou *ordonnée à l'origine* et β_i le coefficient de régression multiple de X_i. Il mesure l'association entre Y et X_i, après avoir tenu compte des associations entre Y et les autres variables. Cette notion convient donc particulièrement bien pour éliminer l'effet des facteurs confondants en épidémiologie.

En pratique, on estime $\beta_0, \beta_1, \cdots, \beta_p$ à partir de données. Les données sont disposées de la manière suivante:

Sujet	Y	X_1	X_2	$\cdots$	X_p
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

On obtient dès lors une estimation de moyenne de la variable Y, elle est notée $\hat{Y}$. C'est la valeur prédite de Y à partir de $\underset{\sim}{X} = \underset{\sim}{x}$.

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

Le résidu $Y_{obs} - \hat{Y}$ ne dépend plus des variables $x_1, \cdots, x_p$.

Test statistique: Pour tester l'effet de la variable X_i sur Y, il suffit de tester l'hypothèse
H_0: $\beta_i = 0$ versus H_1: $\beta_i \neq 0$

On dispose d'une estimation b_i de β_i et de son erreur type $SE(b_i)$.
On peut donc calculer $Z_i = \frac{b_i}{SE(b_i)}$.
Si $|Z_i| > 1.96$, on rejette l'hypothèse H_0.

Exemple
On a mesuré dans un échantillon de données, la PAS, l'âge et le poids et on a modélisé la PAS (Y) en fonction de l'âge (X_1) et du poids (X_2).

On obtient le modèle suivant:

$$\hat{PAS} = 61.8 + 0.653 \times \text{âge} + 0.248 \times poids$$

Le coefficient de régression pour l'âge est positif. Plus l'âge augmente, plus la PAS augmente. Le coefficient de régression pour le poids est positif. Plus le poids augmente, plus la PAS augmente.

Le coefficient de régression pour le poids (0.248) est une mesure de l'association entre le poids et la PAS, qui tient compte de l'association entre l'âge et la PAS.

Remarque
Si on régresse le poids et la PAS en fonction de l'âge, on obtient les équations suivantes:

$$\hat{poids} = 80.9 + 1.869 \times \text{âge}$$

$$\hat{PAS} = 81.9 + 1.116 \times \text{âge}$$

Si pour chacun de ces modèles, on note R_{poids} ($= poids_{obs} - \hat{poids}$) le résidu du poids et R_{PAS} ($= PAS_{obs} - \hat{PAS}$) le résidu de la PAS et que l'on régresse le résidu de la PAS en fonction du résidu du poids, on trouve:

$$R_{PAS} = 0.005 + 0.248 \times R_{poids}$$

Le coefficient de régression du résidu du poids correspond au coefficient de régression du poids dans le modèle de la PAS quand on tient compte de l'âge. Les résidus de la PAS et du poids ont éliminé l'effet de l'âge.

6.3 Régression logistique

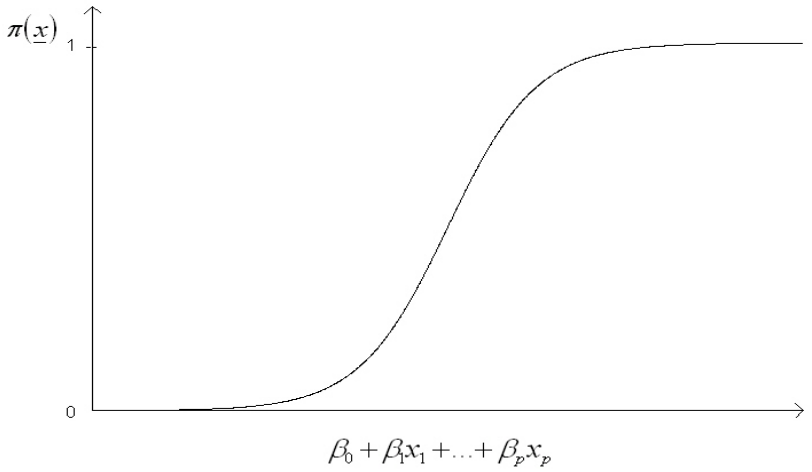
La *régression logistique* consiste à exprimer la relation entre la moyenne d'une variable dépendante binaire (0/1 ou dichotomisée) en fonction de plusieurs variables indépendantes (explicatives).

Y: variable dépendante (binaire / dichotomisée)
$X_1, \cdots, X_p$: variables indépendantes ou explicatives

Puisque Y est une variable binaire, $E\left(Y|\underset{\sim}{x}\right)$ est une proportion notée $\pi(\underset{\sim}{x})$. La régression multiple classique ne convient donc plus car une proportion doit toujours être comprise entre 0 et 1. Cornfield (1967) a défini le *modèle logistique*. Il s'écrit de la façon suivante:

$$\pi(\underset{\sim}{x}) = \frac{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}$$

Il montre que la relation entre la moyenne de Y et les variables $X_1, \cdots, X_p$ n'est plus linéaire mais présente une forme dite logistique (voir figure). De plus, pour tout $\underset{\sim}{x}$, la proportion $\pi(\underset{\sim}{x})$ est toujours comprise entre 0 et 1 comme il se doit.



Le modèle logistique peut s'écrire également:

$$logit\left[\pi(\underset{\sim}{x})\right] = ln\left(\frac{\pi(\underset{\sim}{x})}{1-\pi(\underset{\sim}{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

En pratique, on estime β_0, $\beta_1, \cdots, \beta_p$ à partir de données par la méthode du maximum de vraisemblance. Les données sont disposées de la manière suivante:

Sujet	$Y(0/1)$	X_1	X_2	$\cdots$	X_p
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

On obtient dès lors une estimation de la moyenne de Y, elle est notée $\hat{\pi}$. C'est la valeur prédite de la moyenne de Y à partir de $\underset{\sim}{X} = \underset{\sim}{x}$.

$$ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

Test statistique: Pour tester l'effet de la variable X_i sur Y, il suffit de tester l'hypothèse
H_0: $\beta_i = 0$ versus H_1: $\beta_i \neq 0$

On dispose d'une estimation de b_i de β_i et de son erreur type $SE(b_i)$.
On peut donc calculer $Z_i = \frac{b_i}{SE(b_i)}$.
Si $|Z_i| > 1.96$, on rejette l'hypothèse H_0.

Remarque
On peut démontrer que l'exponentielle du coefficient de régression β_i correspond à l'odds ratio entre Y et X_i corrigé des facteurs confondants. Autrement dit, $e^{\beta_i} = OR_i$ ou $\beta_i = ln(OR_i)$.

En pratique, $b_i = ln(\hat{OR_i})$.
Pour chaque b_i, l'erreur type $SE(b_i)$ est donnée par l'ordinateur. On peut donc calculer un intervalle de confiance à 95% pour chaque β_i, soit

$$b_i - 1.96 \times SE(b_i) \leq \beta_i \leq b_i - 1.96 \times SE(b_i),$$

et donc pour chaque OR_i, en prenant l'exponentielle

$$e^{b_i - 1.96 \times SE(b_i)} \leq e^{\beta_i} \leq e^{b_i + 1.96 \times SE(b_i)}$$

$$\hat{OR}_{i1} \leq OR_i \leq \hat{OR}_{i2}$$

Exemple
Une étude a été réalisée sur la prédiction du développement d'allergies chez le nouveau-né en fonction de l'âge, de la concentration de spermine et de spermidine dans le lait maternel.

Response variable : ALG
Response Levels: 2
Number of observations: 45
Link Function: Logit

Response Profile

Ordered Value	ALG	Count
1	1	14
2	0	31

Analysis of Maximum Likelihood Estimates

Variable		DF	Parameter Estimate b_i	Standard Error $SE(b_i)$	Wald Chi-Square Z_i^2	Pr> Chi-Square p	Odds Ratio
β_0	INTERCPT	1	-1.9159	4.4634	0.1842	0.6677	.
β_1	AGE	1	0.7548	0.8330	0.8210	0.3649	2.127
β_2	SPM	1	-0.9734	0.4321	5.0735	0.0243	0.378
β_3	SPD	1	0.3640	0.3216	1.2810	0.2577	1.439

Pour l'âge, le coefficient de régression vaut 0.7548. L'exponentielle correspond à 2.127 càd à l'estimation de l'odds ratio de l'âge. L'âge n'est pas une variable significative puisque $p > 0.05$ ($Z_i^2 < 3.84$).

L'intervalle de confiance à 95% pour l'odds ratio de l'âge s'obtient en calculant les quantités $e^{0.7548 \pm 1.96 \times 0.8330}$, soit $[0.42 - 10.9]$. Il contient la valeur 1, l'âge n'est donc pas associé à l'allergie.

Si on tient le même raisonnement pour la spermine (SPM), on remarque que c'est une variable significative ($p = 0.0243$). L'intervalle de confiance à 95% pour l'odds ratio est $[0.16 - 0.88]$. La spermine est associée à l'allergie. La probabilité d'être allergique diminue lorsque la concentration de spermine augmente.

Chapitre 7

Valeur diagnostique d'un test

7.1 Introduction

X: Test diagnostique négatif (X_-)
positif (X_+)

M: Maladie à diagnostiquer sur base du test X
M_- : individus ne souffrant pas de la maladie M
M_+ : individus souffrant de la maladie M

$p = P(M)$: *prévalence* de la maladie dans la population. C'est la proportion de sujets atteints de la maladie dans la population.

Le problème est de mesurer l'efficacité du test X à diagnostiquer la maladie M.

Les données sont reprises dans une table 2×2 de la manière suivante (attention à la présentation) qui est différente de celle entre l'association d'une maladie et un facteur de risque):

		Maladie		
		M_-	M_+	
Test diagnostic	X_-	a	b	$a+b$
	X_+	c	d	$c+d$
		$a+c$	$b+d$	n

Il y a deux types d'*échantillonnage* possibles:

- *Séparé*: Echantillon de M_+ et un échantillon de M_- et on observe X_+ et X_-.
- *Mélange*: Echantillon de n personnes dans lequel on observe simultanément X et M.

7.2 Caractéristiques du test

Spécificité du test
Proportion de sujets non-malades pour lesquels le test est négatif.
Proportion de sujets X_- dans M_-
Vrais négatifs (VN)
$SP = \frac{a}{a+c}$

Sensibilité du test
Proportion de sujets malades pour lesquels le test est positif.
Proportion de sujets X_+ dans M_+
Vrais positifs (VP)
$SE = \frac{d}{b+d}$

Faux positifs
Proportion de sujets non-malades pour lesquels le test est positif.
Proportion de sujets X_+ dans M_-
$FP = \frac{c}{a+c} = 1 - SP$

Faux négatifs
Proportion de sujets malades pour lesquels le test est négatif.
Proportion de sujets X_- dans M_+
$FN = \frac{b}{b+d} = 1 - SE$

Efficacité
$EFF = SP + SE - 1$
Pour un test parfait, appelé aussi *pathognomonique* (cfr tuberculose), $SP = SE = 1$

et l'efficacité est aussi égale à 1.

Valeur prédictive positive d'un test

Probabilité d'avoir la maladie quand le test est positif

$$VPP = P\left(M_+|X_+\right) = \frac{pSE}{pSE + (1-p)(1-SP)}$$

NB: Dans le cas où l'échantillonnage se fait du mélange et uniquement dans ce cas, $VPP = \frac{d}{c+d}$.

Valeur prédictive négative d'un test

Probabilité de ne pas avoir la maladie quand le test est négatif

$$VPN = P\left(M_-|X_-\right) = \frac{(1-p)SP}{(1-p)SP + p(1-SE)}$$

NB: Dans le cas où l'échantillonnage se fait du mélange et uniquement dans ce cas, $VPN = \frac{a}{a+b}$.

Exemple

Considérons deux populations, une pour laquelle les individus sont atteints du SIDA (M_+) et une pour laquelle ils n'en sont pas atteints (M_-). Le test diagnostic X à évaluer est le test ELISA.

On a la table suivante:

		SIDA		
		M_-	M_+	
Test ELISA	X_-	987	182	1169
	X_+	13	818	831
		1000	1000	2000

L'échantillonnage n'est pas du mélange mais on a pris 1000 sujets dans chaque cas !

$SP = \frac{987}{1000} = 0.987$
$SE = \frac{818}{1000} = 0.818$
$FP = \frac{13}{1000} = 0.013$
$FN = \frac{182}{1000} = 0.182$
$EFF = 0.987 + 0.818 - 1 = 0.805$

Si la prévalence est de 44 pour 100000, càd $p = 0.00044$,

$VPP = \frac{0.00044 \times 0.818}{0.00044 \times 0.818 + 0.99956 \times 0.013} = 0.028$
$VPN = 0.9999$

Si la prévalence est de 0.10,

$VPP = \frac{0.10 \times 0.818}{0.10 \times 0.818 + 0.90 \times 0.013} = 0.88$
$VPN = 0.9799.$

Chapitre 8

Courbes de survie

Les courbes de survie (*survival analysis*) sont fréquemment utilisées en épidémiologie et dans les essais cliniques.

Références: Kaplan-Meier (1955) et Cox (1972)

8.1 Introduction

La variable étudiée est une durée de vie, notée T. Par exemple, T peut être la durée de vie d'un patient atteint d'un cancer à partir du jour du diagnostic de la maladie.

Propriétés

- T est une variable continue
- $T \geq 0$
- La distribution de T est souvent dissymétrique à droite (mode $<$ médiane $<$ moyenne).
- Certaines données de T peuvent être censurées (*censored observations*). Ce sont des données que l'on ne peut observer mais on sait qu'elles sont supérieures à une certaine valeur. Elles sont souvent affectées d'une astérisque de manière à les identifier. Par exemple, 20^* signifie que le patient a été *perdu de vue* à 20

mois, c'est-à-dire qu'il a été suivi jusqu'au 20^e mois et qu'à ce moment il était toujours en vie. On ne sait pas combien de temps il est resté en vie après le 20^e mois.

Données

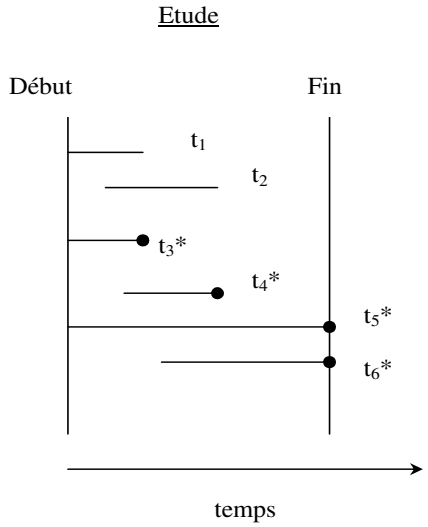
L'échantillon d'effectif n se présente de la façon suivante: $\{t_1, t_2, t_3^*, \cdots, t_n\}$.
Dans un logiciel, on introduit deux variables pour résumer l'échantillon :

Durée de survie: $t_1, t_2, t_3, \cdots, t_n$

Censure: $c_1, c_2, c_3, \cdots, c_n$ où $c_i = 1$ si $t = t_i$ et $c_i = 0$ si $t = t_i^*$

NB: Les censures peuvent être alétoires ou fixes. En général, on postule qu'elles sont aléatoires.

Exemple



Les sujets 1, 3 et 5 entrent dans l'étude au temps 0. Par contre, les sujets 2, 4 et 6 entrent plus tard dans l'étude. Les sujets 1 et 2 décèdent durant l'étude, les temps t_1 et t_2 sont donc des durées de vie réelles. Les sujets 3 et 4 sont perdus de vue pendant l'étude, on sait seulement qu'aux temps t_3 et t_4, ils étaient toujours en vie; les durées de vie t_3 et t_4 sont donc censurées. Les sujets 5 et 6

sont toujours vivants à la fin de l'étude. Les durées de vie t_5 et t_6 sont aussi censurées.

8.2 Courbe de survie

La fonction de survie au temps t est définie comme étant la probabilité de vivre au-delà du temps t.

- $S(t) = P(T > t)$, t allant de 0 à l'infini.
- Par convention, $S(0) = 1$
- $0 \leq S(t) \leq 1$
- La courbe de survie est décroissante
- La durée médiane de survie $t = M$ est la valeur qui correspond à $S(t) = 0.5$.

8.3 Courbe de Kaplan-Meier

La courbe de Kaplan-Meier (KM) est une estimation de la courbe de survie théorique qui utilise toutes les observations y compris les données censurées.

Méthode pour construire la courbe de KM

1. Trier par ordre croissant toutes les données (censurées et non-censurées). En cas d'ex-aequo d'une donnée censurée et d'une donnée non-censurée, la donnée censurée doit toujours suivre la donnée non-censurée.
2. Rechercher les temps non-censurés distincts. Supposons qu'il y en ait k; et notons-les $t_1 < t_2 < \cdots < t_k$
3. Construire une table reprenant pour chaque t_i, l_i qui est le nombre de sujets en vie juste avant t_i et d_i le nombre de sujets qui décèdent en t_i $(i = 1, \cdots, k)$.

4. Calculer pour chaque temps t_i la probabilité de survie $\hat{S}(t_i)$ à l'aide de la formule

$$\hat{S}(t_i) = \prod_{j=1}^{i} \frac{l_j - d_j}{l_j}$$

En clair,
$\hat{S}(t_1) = \frac{l_1 - d_1}{l_1}$
$\hat{S}(t_2) = \frac{l_1 - d_1}{l_1} \times \frac{l_2 - d_2}{l_2}$
$\ldots$
$\hat{S}(t_k) = \hat{S}(t_{k-1}) \times \frac{l_k - d_k}{l_k}$

5. Reporter sur un graphique les valeurs de $\hat{S}(t)$ en fonction des t_i, sachant que l'on commence toujours en 1. On obtient ainsi une courbe en escaliers appelée courbe de Kaplan-Meier.

6. Indiquer les censures par des traits verticaux sur la courbe de survie.

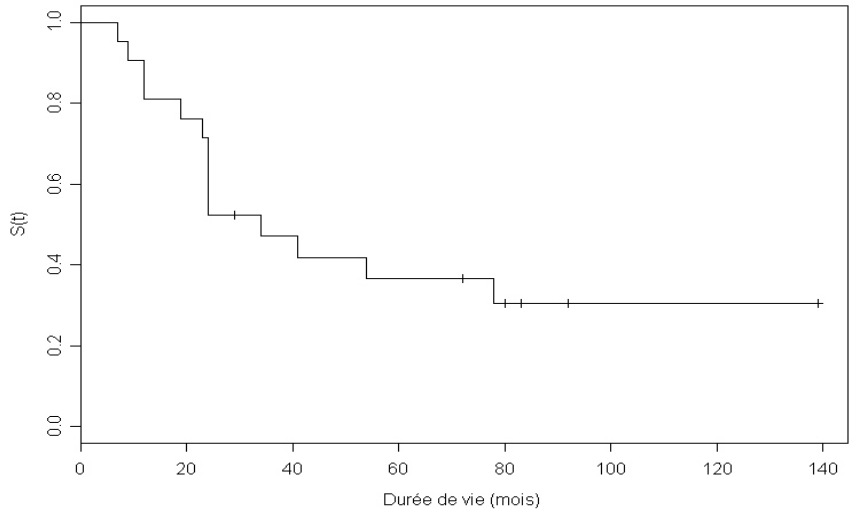
Remarque Si le dernier temps correspond à un décès réel (donnée non-censurée !), la courbe de KM se termine en 0.

Exemple
21 patients atteints d'un cancer colo-rectal ont été traités par radiothérapie. On a mesuré leur durée de survie (mois) après le traitement. On a obtenu les données suivantes (déjà triées par ordre croissant):

$$7, 9, 12, 12, 19, 23, 24, 24, 24, 24, 29^*, 34, 41, 54, 72^*, 78, 80^*, 83^*, 92^*, 139^*, 139^*$$

Notons qu'il y a $k = 10$ temps distincts

t_i	l_i	d_i	$\hat{S}(t_i)$
0	21	0	1
7	21	1	$20/21 = 0.9524$
9	20	1	$0.9524 \times 19/20 = 0.9048$
12	19	2	$0.9048 \times 17/19 = 0.8095$
19	17	1	$0.8095 \times 16/17 = 0.7619$
23	16	1	$0.7619 \times 15/16 = 0.7143$
24	15	4	$0.7143 \times 11/15 = 0.5238$
34	10	1	$0.5238 \times 9/10 = 0.4714$
41	9	1	$0.4714 \times 8/9 = 0.4190$
54	8	1	$0.4190 \times 7/8 = 0.3667$
78	6	1	$0.3667 \times 5/6 = 0.3056$



8.4 Régression de Cox

La méthode de régression de Cox permet d'étudier la relation entre une variable de durée de vie T et des variables indépendantes (ou explicatives) $X_1, \cdots, X_p$.
On a le modèle suivant:

$$S\left(t \left| \underset{\sim}{x} \right.\right) = [S_0(t)]^{e^{\beta_1 x_1 + \cdots + \beta_p x_p}}$$

où $S\left(t \left| \underset{\sim}{x} \right.\right)$ est la courbe de survie des sujets pour lesquels $\underset{\sim}{X} = \underset{\sim}{x}$, $S_0(t)$ une courbe de survie commune à chacun et $\beta_1, \cdots, \beta_p$ les coefficients de régression comme précédemment.

En pratique, il faut estimer les coefficients de régression β_i à partir d'un échantillon de données. Celui-ci se présente comme suit:

Sujet	T	censure	X_1	X_2	$\cdots$	X_p
1	t_1	c_1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	t_2	c_2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n	t_n	c_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

On obtient ainsi une estimation b_i de β_i et son erreur type $SE(b_i)$ comme en régression multiple et en régression logistique.

Pour tester l'hypothèse H_0: $\beta_i = 0$ versus H_1: $\beta_i \neq 0$ (effet ou non de la variable X_i sur la durée de vie), on utilise le test

$$Z_i = \frac{b_i}{SE(b_i)}$$

et on rejette H_0 au niveau d'incertitude de 5% si $|Z_i| > 1.96$. Dans le cas contraire, on peut supposer que $\beta_i = 0$.

Remarque
Considérons le modèle de Cox avec une seule variable indépendante X binaire. On a donc

$$S\left(t \left| \underset{\sim}{x} \right.\right) = [S_0(t)]^{e^{\beta x}}$$

Si $x = 0$, alors $S\left(t \left| \underset{\sim}{0} \right.\right) = [S_0(t)]$
Si $x = 1$, alors $S\left(t \left| \underset{\sim}{1} \right.\right) = [S_0(t)]^{e^{\beta}}$

A partir de cette dernière relation, on peut envisager trois cas possibles:

$\beta = 0$, alors $e^0 = 1$ et $S(t\,|1) = S(t\,|0)$

$\beta > 0$, alors $e^{\beta} > 1$ et $S(t\,|1) < S(t\,|0)$
(il y a donc diminution de la survie, augmentation de risque)

$\beta < 0$, alors $e^{\beta} < 1$ et $S(t\,|1) > S(t\,|0)$
(il y a donc augmentation de la survie, diminution de risque)

Fonction de risque
La fonction $h(t) = \frac{f(t)}{S(t)} = \frac{-dlnS(t)}{dt}$ est appelée la fonction de risque (*hazard function*). C'est le risque instantané de décéder au temps t.

Le modèle de Cox peut alors s'écrire

$$h(t\,|x) = h_0(t).e^{\beta_1 x_1 + \cdots + \beta_p x_p}$$

On comprend dès lors que le risque $h\,(t\,|x\,)$ est proportionnel à $h_0(t)$. C'est pourquoi le modèle de Cox est appelé aussi le modèle des risques proportionnels (*proportional hazard model PH*).

Notons que $h_i = e^{\beta_i}$ est le rapport de risque (*hazard ratio*) qui s'interprète d'une façon fort semblable à l'odds ratio. Une valeur supérieure à 1 correspond à un risque accru et une valeur inférieure à 1 à un risque diminué, une valeur égale à 1 signifiant l'absence de risque pour la variable X_i.

Un intervalle de confiance à 95% peut être obtenu pour h_i en calculant

$b_{i1} = b_i - 1.96 \times SE(b_i)$ et $b_{i2} = b_i + 1.96 \times SE(b_i)$

$\hat{h}_{i1} = e^{b_{i1}}$ et $\hat{h}_{i2} = e^{b_{i2}}$

$\hat{h}_{i1} \leq h_i \leq \hat{h}_{i2}$

Si cet intervalle recouvre la valeur 1, la variable X_i n'est pas un facteur de risque significatif.

Exemple
Une étude a été menée auprès de patients atteints d'un cancer rectal.

Base de données (Fortier et al., 1986)
Obs: Numéro de sujet
T: Durée de vie (mois)
Censure: 0=non, 1=oui
Groupe: Chimiothérapie pré-opératoire (1= $< 5000 rads$, 2= $> 5000 rads$)
Age: Age au moment du diagnostic (années)
Sexe: 0=femme, 1=homme

obs	T	censure	groupe	age	sexe
1	7	0	1	68	0
2	9	0	1	69	0
3	12	0	1	68	0
4	12	0	1	71	0
5	19	0	1	77	1
...	...	...	...	...	...
...	...	...	...	...	...
50	60	1	2	64	0
51	67	0	2	60	1
52	70	0	2	41	1
53	87	1	2	58	1
54	89	1	2	45	1
55	98	1	2	73	1
56	120	1	2	63	1

L'application du modèle des risques proportionnels (ou régression) de Cox sources données conduit aux résultats suivants:

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr> Chi-Square	Hazard Ratio	95% Hazard Confidence	Ratio limits
groupe	1	0.08668	0.49831	0.0303	0.8619	1.091	0.411	2.896
age	1	0.04786	0.02008	5.6818	0.0171	1.049	1.009	1.091
sexe	1	-0.89866	0.47176	3.6287	0.0568	0.407	0.161	1.026

Les variables *groupe* et *sexe* ne sont pas significatives ($p > 0.05$). La variable *age* l'est. Puisque le coefficient de l'âge est positif, cela signifie que les personnes plus âgées présentent un risque accru de décéder. Ceci est confirmé également par le hazard ratio (1.049) qui est supérieur à 1.

Chapitre 9

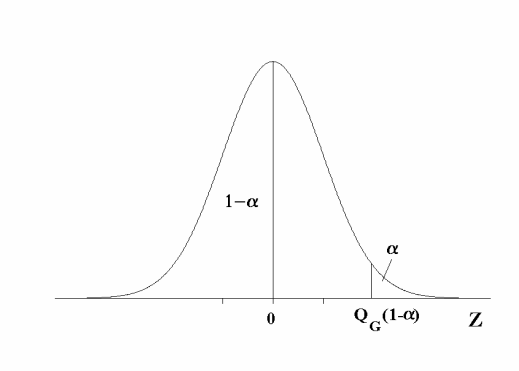
Annexes

9.1 Table de nombres aléatoires

10	9	73	25	33	76	52	1	35	86	34	67	35	48	76	80	95	90	91	17
37	54	20	48	5	64	89	47	42	96	24	80	52	40	37	20	63	61	4	2
8	42	26	89	53	19	64	50	93	3	23	20	90	25	60	15	95	33	47	64
99	1	90	25	29	9	37	67	7	15	38	31	13	11	65	88	67	67	43	97
12	80	79	99	70	80	15	73	61	47	64	3	23	66	53	98	95	11	68	77
66	6	57	47	17	34	7	27	68	50	36	69	73	61	70	65	81	33	98	85
31	6	1	8	5	45	57	18	24	6	35	30	34	26	14	86	79	90	74	39
85	26	97	76	2	2	5	16	56	92	68	66	57	48	18	73	5	38	52	47
63	57	33	21	35	5	32	54	70	48	90	55	35	75	48	28	16	82	87	9
73	79	64	57	53	3	52	96	47	78	35	80	83	42	82	60	93	52	3	44
98	52	1	77	67	14	90	56	86	7	22	10	94	5	58	60	97	9	34	33
11	80	50	54	31	33	80	82	77	32	50	72	56	82	48	29	40	52	42	1
83	45	29	96	34	6	28	89	80	83	13	74	67	0	78	18	47	54	6	10
88	68	54	2	0	86	50	75	84	1	36	76	66	79	51	90	36	47	64	93
99	59	46	73	48	87	51	76	49	69	91	82	60	89	28	93	78	56	13	68
65	48	11	76	74	17	46	85	9	50	58	4	77	69	74	73	3	95	71	86
80	12	43	56	35	17	72	70	80	15	45	31	82	23	74	21	11	57	82	53
74	35	9	98	17	77	40	27	72	14	43	23	60	2	10	45	52	16	42	37
69	91	62	68	3	66	25	22	91	48	36	93	68	72	3	76	62	11	39	90
9	89	32	5	5	14	22	56	85	14	46	42	75	67	88	96	29	77	88	22
91	49	91	45	23	68	47	92	76	86	46	16	28	35	54	94	75	8	99	23
80	33	69	45	98	26	94	3	68	58	70	29	73	41	35	53	14	3	33	40
44	10	48	19	49	85	15	74	79	54	32	97	92	65	75	57	60	4	8	81
12	55	7	37	42	11	10	0	20	40	12	86	7	46	97	96	64	48	94	39
63	60	64	93	29	16	50	53	44	84	40	21	95	25	63	43	65	17	70	82
61	19	69	4	46	26	45	74	77	74	51	92	43	37	29	65	39	45	95	93
15	47	44	52	66	95	27	7	99	53	59	36	78	38	48	82	39	61	1	18
94	55	72	85	73	67	89	75	43	87	54	62	24	44	31	91	19	4	25	92
42	48	11	62	13	97	34	40	87	21	16	86	84	87	67	3	7	11	20	59
23	52	37	83	17	73	20	88	98	37	68	93	59	14	16	26	25	22	96	63
4	49	35	24	94	75	24	63	38	24	45	86	25	10	25	61	96	27	93	35
0	54	99	76	54	64	5	18	81	59	96	11	96	38	96	54	69	28	23	91
35	96	31	53	7	26	89	80	93	54	33	35	13	54	62	77	97	45	0	24
59	80	80	83	91	45	42	72	68	42	83	60	94	97	0	13	2	12	48	92
46	5	88	52	36	1	39	9	22	86	77	28	14	40	77	93	91	8	36	47

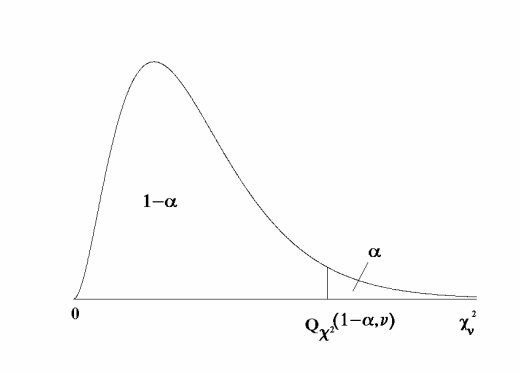
From tables of the RAND Corporation, by permission.

9.2 Quantiles de la loi Normale $N(0,1)$



Aire supérieure (α)	Aire inférieure ($1-\alpha$)	Quantile $Q_G(1-\alpha)$
1.0	0	$-\infty$
0.995	0.005	-2.58
0.99	0.01	-2.33
0.975	0.025	-1.96
0.95	0.05	-1.645
0.5	0.5	0
0.05	0.95	1.645
0.025	0.975	1.96
0.01	0.99	2.33
0.005	0.995	2.58
0.0	1.0	$+\infty$

9.3 Quantiles de la loi Chi-carré à ν degrés de liberté



Degrés de liberté ν	Quantile 0.95	Quantile 0.99
1	3.84	6.64
2	5.99	9.21
3	7.82	11.3
4	9.49	13.3
5	11.1	15.9
6	12.6	16.8
7	14.1	18.5
8	15.6	20.1
9	16.9	21.7
10	18.3	23.2

Chapitre 10

Exercices

10.1 Introduction

Exercice 1
Une population contient $N = 5$ sujets dont les poids (variable X) valent respectivement 48, 49, 53, 55 et 60 kg.

1. Calculez la moyenne μ et la variance σ^2 de cette population
2. Calculez la moyenne $\overline{x}$ de tous les échantillons d'effectif $n = 2$ (avec remplacement) extraits de cette population
3. Calculez la moyenne de ces moyennes et montrez qu'elle est égale à μ
4. Calculez la variance de ces moyennes et montrez qu'elle est égale à $\frac{\sigma^2}{n}$
5. Refaites les calculs des points (c) et (d) en vous limitant aux échantillons sans remplacement (l'ordre n'a pas d'importance). Montrez que la moyenne vaut toujours μ et que la variance vaut $\frac{\sigma^2}{n}\frac{N-n}{N-1}$.

Exercice 2
Dans une population de 500 sujets d'âge moyen 20 ans et d'écart-type 3 ans, on tire des échantillons d'effectif 50. Que valent la moyenne et la variance d'échantillonnage des moyennes d'échantillons?

1. Utilisez la formule générale
2. Utilisez la formule simplifiée et comparez.

Exercice 3
La prévalence d'une maladie dans une population de 1000 individus vaut $\pi = 15\%$. Si on tire des échantillons d'effectifs $n = 200$ de cette population et que l'on calcule la proportion p, que vaut l'erreur type de p?

Exercice 4
Un sondage d'opinion portant sur 160 personnes d'une population en comportant 1200 a révélé que 35% fumaient.

1. Quelle est la précision (erreur type) de cette estimation?
2. Calculez un intervalle de confiance (fourchette) à 95% pour la proportion réelle π
3. Il y a 10 ans dans la même population, 48% des sujets fumaient. La diminution est-elle significative?

Exercice 5
Dans une population très grande, on a extrait un échantillon de 100 personnes. Le taux de cholestérol moyen observé est de 2,5 g/l avec un écart-type de 0,75 g/l.

1. Calculez l'erreur type de la moyenne
2. Déterminez un intervalle de confiance à 95 pour la vraie moyenne μ

Exercice 6
Dans une étude épidémiologique impliquant 20 enquêteurs bilingues, dix couvriront la Flandre (groupe A) et dix la Wallonie et Bruxelles (groupe B). En vous servant de vos tables de nombres aléatoires, établissez la liste des enquêteurs du groupe A et du groupe B. Pour l'ensemble de la classe, répertoriez l'attribution de l'enquêteur N°8 dans les deux groupes!

Exercice 7
Dans un groupe de 50 sujets, 10 sont séropositifs.

1. Que vaut la moyenne μ ($= \pi$) et la variance σ^2 de la variable binaire "séropositivité" (0=non, 1=oui)?
2. Que valent la moyenne et l'erreur type des échantillons d'effectif $n = 5$ extraits de cette population (sans remplacement)?

10.2 Echantillonnage

Exercice 1
Dans une population composée de trois strates comportant chacune 100 individus, les proportions de sujets présentant une anomalie cardiaque valent respectivement 5%, 15% et 50%. On décide de tirer 30 sujets de la population.

1. Combien doit-on tirer de sujets de chaque strate pour obtenir un échantillonnage optimum?

2. Dans le cas d'un échantillonnage simplement fortuit, que valent la moyenne et la variance d'échantillonnage de la proportion d'échantillon p?

3. Dans le cas d'un échantillonnage stratifié d'effectif 10 de chaque strate, que valent la moyenne et la variance d'échantillonnage de la proportion p globale?

4. Y a-t-il réduction de la variabilité d'échantillonnage dans le cas 3 par rapport à 2?

Exercice 2
Quand l'échantillonnage stratifié conduit-il à une variance d'échantillonnage plus faible? Quand l'échantillonnage stratifié est-il contre-indiqué?

Exercice 3
Voici une population de $N = 20$ sujets atteints (1) ou non (0) d'infarctus du myocarde

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	1	0	0	1	0	0	0	1	1	0	1	0	1	0	0	1	1	0	1

Envisagez tous les échantillons systématiques d'effectif n=4 de cette population

1. Combien y en a-t-il?

2. Calculez chaque fois la moyenne p

3. Que vaut la moyenne des p?

4. Que vaut la variance des p?

5. Qu'aurait valu la moyenne et la variance d'échantillonnage, si on avait extrait un échantillonnage simplement fortuit d'effectif n=4?

Exercice 4

Considérons trois grappes (G1, G2, G3) contenant chacune 4 sujets, fumeurs ou non fumeurs (X = 0 ou 1) comme suit:

G1				G2				G3			
A	B	C	D	E	F	G	H	I	J	K	L
1	0	0	0	0	1	1	0	0	1	1	1

1. Quelle est la variabilité d'échantillonnage des proportions d'échantillons *simplement fortuit* d'effectif $n = 3$ extraits de cette population?

2. Quelle est la variabilité d'échantillonnage des proportions d'échantillons *stratifiés* d'effectif $n = 3$ extraits de cette population?

3. Quelle est la variabilité d'échantillonnage des proportions d'échantillons d'une *grappe* d'effectif $n = 3$ extraits de cette population?

10.3 Taille d'échantillon

Exercice 1

On estime que la proportion réelle π d'une anomalie génétique donnée dans une population se situe entre 0.05 et 0.10. Quel doit être l'effectif n d'un échantillon simplement fortuit extrait de cette population pour que la proportion observée ne s'écarte pas plus de 0.005 de la même valeur au niveau d'incertitude de 5%? Faites varier π de 0.05 à 0.10 par pas de 0.01.

Exercice 2

En reprenant l'énoncé de l'exercice 1, quel doit être l'effectif n d'un échantillon simplement fortuit extrait de cette population pour que la proportion observée ne s'écarte pas plus de 10% de la même valeur ($\alpha = 5\%$)? Faites varier π de 0.05 à 0.10 par pas de 0.01.

Exercice 3

Dans une population finie d'effectif $N = 1000$, 75% des sujets fument et la pression artérielle systolique moyenne vaut 165 mmHg avec un écart-type de 15 mmHg. On tire un échantillon simplement fortuit d'effectif n de cette population. On fixe $\alpha = 5\%$

1. Quelle doit être la valeur n pour que la proportion de fumeurs observée ne s'écarte pas plus de 10% de la valeur théorique?

2. Pour quelle valeur de n, la pression artérielle systolique moyenne ne s'écarte-t-elle pas de plus d'un tiers d'écart-type théorique de la vraie moyenne de population?

3. Trouvez la valeur de n pour que la proportion de fumeurs observée et la moyenne de PAS ne s'écartent chacune pas plus de 15% de leur valeur réelle?

Exercice 4

Quel doit être l'effectif n d'un échantillon simplement fortuit extrait d'une population finie d'effectif $N = 200$ et de variance $\sigma^2 = 10$ pour que la variance d'échantillonnage des moyennes d'échantillon soit égale à 0,45?

Exercice 5

Le taux d'incidence sur 10 ans d'une affection cardio-vasculaire est estimé à 15% en Belgique et à 25% en France. On souhaite échantillonner les populations de ces deux pays dans un rapport 1:4 (un belge pour quatre français). Déterminez la taille des échantillons nécessaires pour mettre en évidence (au niveau d'incertitude $\alpha = 5\%$ et avec un risque de seconde espèce $\beta = 10\%$) la différence entre les taux d'incidence des deux pays?

Exercice 6

Dans le même contexte (voir exercice 5), le taux moyen de cholestérol vaut 1,66 mmol/L en Belgique et 1,78 mmol/L en France (avec une même dispersion $\sigma = 0,85$ mmol/L). Quels doivent être les effectifs des échantillons (n et $4n$) pour mettre en évidence la différence entre les taux moyens de cholestérol?

Exercice 7

On veut tester l'hypothèse nulle que le taux de mortalité dans une population de malades est de 20%. On souhaite que si le taux de mortalité est en réalité de 25%, on puisse le mettre en évidence à partir d'un échantillon d'effectif n extrait de la population. Que vaut la valeur de n, si $\alpha = 0.05$ et $\beta = 0.20$?

10.4 Risque relatif

Exercice 1

La table ci-dessous montre l'association entre le taux de cholestérol et le taux d'incidence de maladie cardio-vasculaire (sur une période de 18 ans).

Maladie cardio-vasculaire

Cholestérol	Oui	Non	Total à risque
< 260	177	800	977
≥ 260	91	295	386
Total	268	1095	1363

1. Calculez le risque relatif (RR) d'acquérir la maladie en fonction du taux de cholestérol.

2. Calculez l'intervalle de confiance à 95% pour RR en utilisant la méthode de Katz.

Exercice 2

La table ci-dessous est sortie d'un rapport concernant la relation entre la prise d'aspirine et les attaques cardiaques (Harvard Medical School).

Infarctus du myocarde

Traitement	Oui	Non	Total à risque
Placebo	189	10845	11034
Aspirine	104	10933	11037
Total	293	21778	22071

1. Calculez le risque relatif (RR) d'avoir un infarctus en fonction du traitement.

2. Calculez l'intervalle de confiance à 95% pour RR en utilisant la méthode de Katz.

Exercice 3

La table ci-dessous montre l'association entre le port de la ceinture de sécurité et le taux de mortalité sur base sur des accidents de la route survenus en Floride en 1988.

Blessures mortelles

Ceinture	Oui	Non	Total à risque
Non	1601	162527	164128
Oui	510	412368	412878
Total	2111	574895	577006

1. Calculez le risque relatif (RR) de décéder en fonction du port de la ceinture de sécurité.

2. Calculez l'intervalle de confiance à 95% pour RR en utilisant la méthode de Katz.

10.5 Odds ratio

Exercice 1

La table ci-dessous montre l'association entre le taux de cholestérol et le taux d'incidence de maladie cardio-vasculaire (sur une période de 18 ans).

Maladie cardio-vasculaire

Cholestérol	Oui	Non	Total à risque
< 260	177	800	977
≥ 260	91	295	386
Total	268	1095	1363

1. Calculez l'odds ratio (OR) et l'intervalle de confiance à 95% par la méthode de Woolf.

2. Calculez le risque attribuable dans le groupe exposé (AR) et dans la population (AR_p).

Exercice 2

La table ci-dessous est sortie d'un rapport concernant la relation entre la prise d'aspirine et les attaques cardiaques (Harvard Medical School).

Infarctus du myocarde

Traitement	Oui	Non	Total à risque
Placebo	189	10845	11034
Aspirine	104	10933	11037
Total	293	21778	22071

1. Calculez l'odds ratio (OR) et l'intervalle de confiance à 95% par la méthode de Woolf.

2. Calculez le risque attribuable dans le groupe exposé (AR) et dans la population (AR_p).

Exercice 3

La table ci-dessous montre l'association entre le port de la ceinture de sécurité et le taux de mortalité sur base sur des accidents de la route survenus en Floride en 1988.

Blessures mortelles

Ceinture	Oui	Non	Total à risque
Non	1601	162527	164128
Oui	510	412368	412878
Total	2111	574895	577006

1. Calculez l'odds ratio (OR) et l'intervalle de confiance à 95% par la méthode de Woolf.

2. Calculez le risque attribuable dans le groupe exposé (AR) et dans la population (AR_p).

10.6 Facteur confondant

Exercice 1

Une étude visant à mesurer l'association entre une maladie M et l'exposition E à une substance toxique a conduit aux deux tables suivantes en fonction du sexe (supposé être un facteur confondant).

Hommes

	M_+	M_-	
E_+	9100	990	10090
E_-	900	8110	9010
Total	10000	9100	19100

Femmes

	M_+	M_-	
E_+	900	10	910
E_-	89100	81890	170990
Total	89100	81890	170990

1. Calculez l'odds ratio (OR) dans chaque groupe.
2. Calculez l'odds ratio de Mantel-Haenszel (OR_{MH}).
3. Calculez l'odds ratio de Woolf (OR_W).
4. Effectuez un test d'interaction ($OR_H = OR_F$).

Exercice 2

Sur base d'une étude de prévalence dans des populations à statut socio-économique (SSE) élevé ou bas, on a obtenu les tables d'association entre le cancer du sein (K) et la prise de réserpine (R).

SSE élevé

	M_+	M_-	
E_+	16	79984	80000
E_-	184	919816	920000
Total	200	999800	1000000

SSE bas

	M_+	M_-	
E_+	4	19996	80000
E_-	196	979804	920000
Total	200	999800	1000000

1. Calculez l'odds ratio (OR) dans chaque groupe.
2. Calculez l'odds ratio de Mantel-Haenszel (OR_{MH}).
3. Calculez l'odds ratio de Woolf (OR_W).

4. Effectuer un test d'interaction ($OR_{SSEeleve} = OR_{SSEbas}$).

Exercice 3

Voici les tables d'association entre le développement d'une maladie cardio-vasculaire (MCV) et la pression artérielle systolique (PAS) en fonction de trois catégories d'âge chez l'homme.

Hommes 45-49 ans

	M_+	M_-	
E_+	9	17	26
E_-	36	147	183
Total	45	164	209

Hommes 50-54 ans

	M_+	M_-	
E_+	14	21	35
E_-	35	131	166
Total	49	152	201

Hommes 55-59 ans

	M_+	M_-	
E_+	22	28	50
E_-	45	127	172
Total	67	155	222

1. Calculez l'odds ratio (OR) dans chaque groupe.
2. Calculez l'odds ratio de Mantel-Haenszel (OR_{MH}).
3. Calculez l'odds ratio de Woolf (OR_W).
4. Effectuer un test d'interaction ($OR_{Hommes45-49} = OR_{Hommes50-54} = OR_{Hommes55-59}$).

10.7 Régression multiple

On a étudié la dépendance entre le Citalopram (anti-dépresseur) et d'autres types de médicaments. Pour cela, on a mesuré chez 30 patients le taux de Citalopram dans le sang, la dose donnée (*DOSE*), le fait que le patient prenait ou non des neuroleptiques (*NEUROLEP*), des vitamines (*VITAMINE*) ou d'autres médicaments (*DIVERS*). Pour normaliser la distribution, on a utilisé le logarithme du taux de citalopram (*LTAUX*).

1. Nombre d'observations ?
2. Quelle est la variable dépendante et quelles sont les variables explicatives.
3. Quelle est l'équation du modèle ?
4. Les variables explicatives sont-elles significatives ? Calculez le Z.
5. Interprétez les résultats
6. Sachant que Mme Durant prend 50 gr de citalopram et qu'elle ne prend que des vitamines, quel pourrait être son taux de citalopram dans le sang?

Model: MODEL1
Dependent Variable: LTAUX

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	4	12.89331	3.22333	9.055	0.0001
Error	25	8.89902	0.35596		
C Total	29	21.79233			

Root MSE	0.59662	R-square	0.5916
Dep Mean	3.49181	Adj R-sq	0.5263
C.V.	17.08639		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	2.093912	0.37774031		0.0001
DOSE	1	0.054184	0.01252206		0.0002
NEUROLEP	1	0.506899	0.23038442		0.0373
VITAMINE	1	-0.528653	0.28129546		0.0719
DIVERS	1	-0.442101	0.23562156		0.0723

10.8 Régression logistique

Une étude a été réalisée au CHU en cardiologie sur les complications qu'il pouvait y avoir après une coronarographie. On s'est intéressé aux complications à l'hôpital oui/non (*COMPLHC*), au sexe (*SEXE* 1=M, 0=F), à l'âge (*AGE*), à l'angor stable oui/non (*ANGORST*), angor instable oui/non (*ANGORIN*), stent oui/non (*STENT*) et hypertension oui/non (*HTA*).

1. Nombre d'observations ?
2. Quelle est la variable dépendante et quelles sont les variables explicatives.
3. Quelle est l'équation du modèle ?
4. Calculez les OR pour chaque variable ainsi que les intervalles de confiance de ces OR
5. Les variables explicatives sont-elles significatives ? Calculez le Z^2 (=Wald chi-square).
6. Interprétez les résultats
7. Sachant que Mr Dupont est âgé de 75 ans, qu'il n'a ni angor stable ni angor instable ni hypertension mais qu'il a un stent, quelle est la probabilité qu'il ne fasse pas de complications à l'hôpital ?

The LOGISTIC Procedure

Data Set: WORK.PAC
Response Variable: COMPLHC
Response Levels: 2
Number of Observations: 2002
Link Function: Logit

Response Profile

Ordered Value	COMPLHC	Count
1	1	112
2	0	1890

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	865.496	847.767	.
SC	871.098	886.980	.
-2 LOG L	863.496	833.767	29.729 with 6 DF (p=0.0001)
Score	.	.	30.231 with 6 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Odds Ratio
INTERCPT	1	-5.9277	0.7604		0.0001	
SEXE	1	0.1831	0.2386		0.4427	
AGE	1	0.0402	0.0102		0.0001	
ANGORST	1	0.0713	0.2861		0.8032	
ANGORIN	1	0.4793	0.2417		0.0474	
STENT	1	0.6402	0.2218		0.0039	
HTA	1	0.0911	0.2000		0.6487	

Avec les mêmes données que dans l'exercice précédent, on a modélisé par un modèle de régression logistique les complications à l'hôpital en fonction du fait d'avoir ou non un stent.

1. Montrez qu'on peut calculer l'odds ratio de deux façons différentes.
2. Calculez l'intervalle de confiance à 95% de l'OR.
3. Interprétez le modèle

		COMPLHC		
	Freq	0	1	Total
Stent	0	1559	80	1639
	1	331	32	363
	Total	1890	112	2002

Data Set: WORK.PAC
Response Variable: COMPLHC
Response Levels: 2
Number of Observations: 2002
Link Function: Logit

Response Profile

Ordered Value	COMPLHC	Count
1	1	112
2	0	1890

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	865.496	859.727	.
SC	871.098	870.931	.
-2 LOG L	863.496	855.727	7.769 with 1 DF (p=0.0053)
Score	.	.	8.710 with 1 DF (p=0.0032)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Odds Ratio
INTERCPT	1	-2.9698	0.1146		0.0001	
STENT	1	0.6334	0.2177		0.0036	

10.9 Valeur diagnostique d'un test

Exercice 1

Un total de 9863 sujets d'une communauté A ont été classés en fonction de la présence ou de l'absence d'arthrite rhumatoïde et du résultat (négatif ou positif) d'un nouveau test diagnostique de cette maladie. Voici les résultats obtenus.

		Arthrite rhumatoïde		
		Absente	Présente	
Test	Négatif	7989	99	8088
	Positif	888	887	1775
	Total	8877	986	9863

1. Calculez la spécificité, la sensibilité et l'efficacité du nouveau test avec intervalle de confiance à 95%

2. Que vaut la prévalence de la maladie $P(M)$?

3. Calculez la valeur prédictive positive (VPP) enégative (VPN) du test en utilisant la prévalence obtenue au point 2.

4. Refaire le point 3 en utilisant les formules simplifiées. Que constatez-vous ?

Exercice 2

Sur 5550 sujets d'une communauté B, on a par contre obtenu la table suivante :

		Arthrite rhumatoïde		
		Absente	Présente	
Test	Négatif	3465	165	3630
	Positif	385	1485	1870
		3850	1650	5500

Répondre aux mêmes questions (a-d) qu'à l'exercice 1.

Exercice 3

Le pic de créatinine kérase (CK) est souvent considéré comme un bon test diagnostic de l'infarctus myocardique aigu (IMA). On a évalué ce test chez 500 patients sans

IMA et chez 400 patients atteints d'un IMA. Voici les résultats.

		Infarctus myocardique aigu		
		Absente	Présente	
PIC CK	$\leq$ 100 UI/C Négatif	454	13	467
	$>$ 100 UI/C Positif	46	387	433
		500	400	900

1. Calculer la spécificité et la sensibilité du test (avec intervalle de confiance à 95%).

2. Calculer la VPP du test pour $P(M)$=0.05, 0.10 et 0.20.

10.10 Courbes de Kaplan-Meier

Exercice 1

Calculez les courbes de survie de Kaplan-Meier suivantes et faites chaque fois un graphique. Quelle est la durée médiane de survie dans chaque cas (graphiquement) ?

1. 5, 5*, 6, 7, 8, 8, 8*, 10, 11*, 25, 26*
2. 7*, 7*, 8*, 9, 9, 10, 11, 11, 18, 18*, 19, 20, 20*, 27
3. 10, 11*, 12, 16, 16*, 20, 21, 22, 30, 32*, 34*, 42*, 50*
4. 12, 7, 12, 15, 14*, 17, 6*, 20*, 24, 23, 32*, 36, 37, 40
5. 10, 10, 10, 10, 10, 15, 20, 20*, 30, 30, 35, 35, 40*, 50, 50
6. 100, 106, 110, 110, 120, 50, 45, 30, 30, 110, 120, 140
7. 20, 30, 30, 40, 50, 60, 70, 20*, 20*, 30*, 60*, 70*, 80*
8. 34, 21, 21*, 21, 33, 18, 21, 22*, 24, 10*, 10*, 8*, 7, 21

Exercice 2

Freirech et al ont réalisé une étude sur la leucémie chez 42 enfants : 21 traités avec un placebo et 21 traités avec de la 6-mercaptopurine (6-MP). Les patients ont été suivis jusqu'à ce qu'ils aient une rechute ou jusque la fin de l'étude (auquel cas, ils sont censurés). Les durées sont données en mois.

Voici les résultats :

Placebo : 1, 22, 3, 12, 8, 17, 2, 11, 8, 12, 2, 5, 4, 15, 8, 23, 5, 11, 4, 1, 8

6-MP : 10, 7, 32*, 23, 22, 6, 16, 34*, 32*, 25*, 11*, 20*, 19*, 6, 17*, 35*, 6, 13, 9*, 6*, 10*

1. Calculez les courbes de survie (rechute) de Kaplan-Meier pour chaque groupe et représentez les sur le même graphique.

2. Quelle est la durée médiane de rechute dans chaque groupe ?

3. A partir du graphique, dans quel groupe peut-on dire qu'il y a le plus grand risque de rechute ?

10.11 Régression de Cox

On a étudié la durée de survie après une greffe de moëlle osseuse chez des patients ayant la leucémie. Pour cela, on a mesuré chez 4381 patients la durée de survie (*YRSUR*, années), la censure de survie (*ETASUR* 1=en vie, 2=décédé), le sexe (*SX*, 1=homme, 0=femme), l'âge (*AGECAT1*=1 si âge entre 18 et 40 ans et 0 sinon, *AGECAT2*=1 si âge ¿ 40 ans et 0 sinon) et le type de greffe (*TYPM1*=1 si autogreffe et 0 sinon, *TYPM2*=1 si allogreffe jumeau/fratrie et 0 sinon, *TYPM3*=1 si allogreffe non famille et 0 sinon). A partir de la régression de Cox ci-dessous,

1. Quel est le nombre d'observations utilisées ?
2. Combien de personnes sont décédées ?
3. Quelles variables sont significatives (?=0.05) ?
4. Donnez une interprétation du modèle.

The PHREG Procedure

Informations sur le modèle

Data Set	WORK.TOUS	
Dependent Variable	yrsur	
Censoring Variable	ETASUR	ETASUR
Censoring Value(s)	1	
Ties Handling	EXACT	

Number of Observations Read	4381
Number of Observations Used	4361

Récapitulatif du nombre d'événements et de valeurs tronquées

Total	Événement	Tronqué (e)	Pourcentage tronqué
4361	1724	2637	60.47

État de convergence

Convergence criterion (GCONV=1E-8) satisfied.

Statistiques d'ajustement du modèle

Critère	Sans covariables	Avec covariables
-2 LOG L	24450.795	24240.320
AIC	24450.795	24252.320
SBC	24450.795	24285.034

Test de l'hypothèse nulle globale : BETA=0

Test	Khi 2	DF	Pr > Khi 2
Likelihood Ratio	210.4752	6	<.0001
Score	268.4081	6	<.0001
Wald	251.7178	6	<.0001

Analyse des estimations de la vraisemblance maximum

Variable	DF	Résultat estimé des paramètres	Erreur std	Khi 2	Pr > Khi 2	Rapport de risque	95% Limites de confiance du rapport de risque	
sx	1	-0.05014	0.04890	1.0514	0.3052	0.951	0.864	1.047
agct1	1	0.34195	0.07810	19.1704	<.0001	1.408	1.208	1.641
agct2	1	0.38964	0.07866	24.5380	<.0001	1.476	1.266	1.723
typm1	1	0.32064	0.06062	27.9794	<.0001	1.378	1.224	1.552
typm2	1	0.81696	0.13020	39.3705	<.0001	2.264	1.754	2.922
typm3	1	1.17501	0.07967	217.5146	<.0001	3.238	2.770	3.785