Guide d'utilisation d'Epi Info 7 pour réaliser des analyses statistiques

Loic Desquilbet

Guide d'utilisation d'Epi Info 7 pour réaliser des analyses statistiques

Loïc Desquilbet, PhD en Santé Publique

Professeur en Biostatistique et en Epidémiologie Clinique
Département des Sciences Biologiques et Pharmaceutiques
Ecole nationale vétérinaire d'Alfort

Version 1.0 (24/01/2022)

Préface

Comment lire ce guide

Chaque partie de ce guide d'utilisation du logiciel Epi Info 7 ne peut pas se lire avant d'avoir lu les parties précédentes. Ainsi, si par exemple vous souhaitez utiliser un modèle de Cox pour analyser vos données, vous devrez lire … l'intégralité de ce guide !

Contrat de diffusion

Citation du document

Vous pouvez citer ce document de la façon suivante : Desquilbet, L. 2022. *Guide d'utilisation d'Epi Info 7 pour réaliser des analyses statistiques*. [En ligne, disponible à : https://hal.archives-ouvertes.fr/hal-*numéro à modifier*]

Suggestions de modifications de ce document

Si vous avez des suggestions de modifications de ce guide (coquilles dans le texte, souhaits de clarification de certains passages, etc.), n'hésitez pas à me les signaler par email (loic.desquilbet(at)vet-alfort.fr). Je vous remercie beaucoup par avance.

Table des matières

Chapitre 3 - Modèles de régression

Chapitre 1 - Préambule

I. Présentation rapide du logiciel Epi Info 7

A. Téléchargement d'Epi Info 7 et comparaison avec d'autres logiciels

Epi Info 7 est un logiciel gratuit qui ne fonctionne (malheureusement) que sur PC. Il peut se télécharger en cliquant ici. Ce logiciel permet entres autres de concevoir des questionnaires, de saisir des données, et de réaliser des analyses statistiques sur des données. C'est cette dernière fonctionnalité du logiciel que traite ce guide, et ce, de façon non exhaustive. L'aide (en anglais) du logiciel dans sa partie « analyse des données » se trouve ici.

Pour quelle raison utiliser Epi Info et non pas d'autres logiciels de statistique ? Tout d'abord, citons quelques logiciels de statistique que l'on peut utiliser, dont certains sont payants, et d'autres gratuits(*) : BiostaTGV*, R*, GraphPad, XLSTAT, et SAS*[1]. Les avantages d'Epi Info par rapport aux trois logiciels gratuits précédemment cités sont les suivants. Par rapport à BiostaTGV, Epi Info travaille directement sur un fichier de données importé d'Excel par exemple et il permet de réaliser des modèles de régression (linéaire, logistique, et Cox). Par rapport à R et SAS, Epi Info est beaucoup plus simple d'utilisation, et ne demande aucune connaissance / appétence en langage de programmation. Le gros inconvénient d'Epi Info est qu'il ne fonctionne que sur PC, et pas sur Mac. Il est aussi bien plus limité que ne le sont les autres logiciels de statistique dans les différentes méthodes d'analyses statistiques proposées.

B. Premier lancement d'Epi Info 7

Après avoir installé Epi Info, en lançant Epi Info, vous obtenez la fenêtre ci-dessous (cf. Figure 1). La toute première démarche que je vous suggère (et que vous n'aurez à faire qu'une seule fois) est de choisir l'anglais comme langue par défaut (si tel n'est pas déjà le cas juste après le téléchargement). Ainsi, les analyses statistiques décrites dans Epi Info correspondront à celles que vous lisez dans la littérature scientifique (majoritairement anglophone).

[1] Gratuit pour les universitaires et les étudiants

Figure 1

Pour cela, vous cliquez sur « Créer questionnaires » (Figure 1.a), puis sur « Outils » → « Options » → « Langue » et on sélectionne « English (default) ». Puis vous redémarrez Epi Info pour prendre en compte le changement de langue. En relançant Epi Info, vous obtenez la fenêtre ci-dessous (Figure 2).

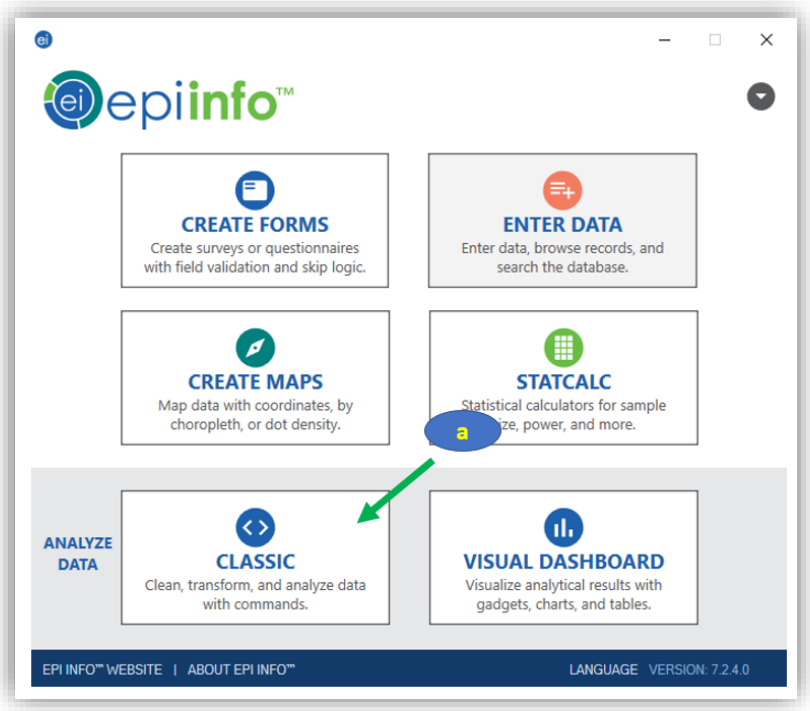


Figure 2

Vous cliquez sur « Classic » dans « Analyse data » (Figure 2.a), car c'est en effet la seule fenêtre que je vais utiliser dans ce guide (à part au moment de créer un projet pour enregistrer un programme). On obtient alors la fenêtre ci-dessous (Figure 3).

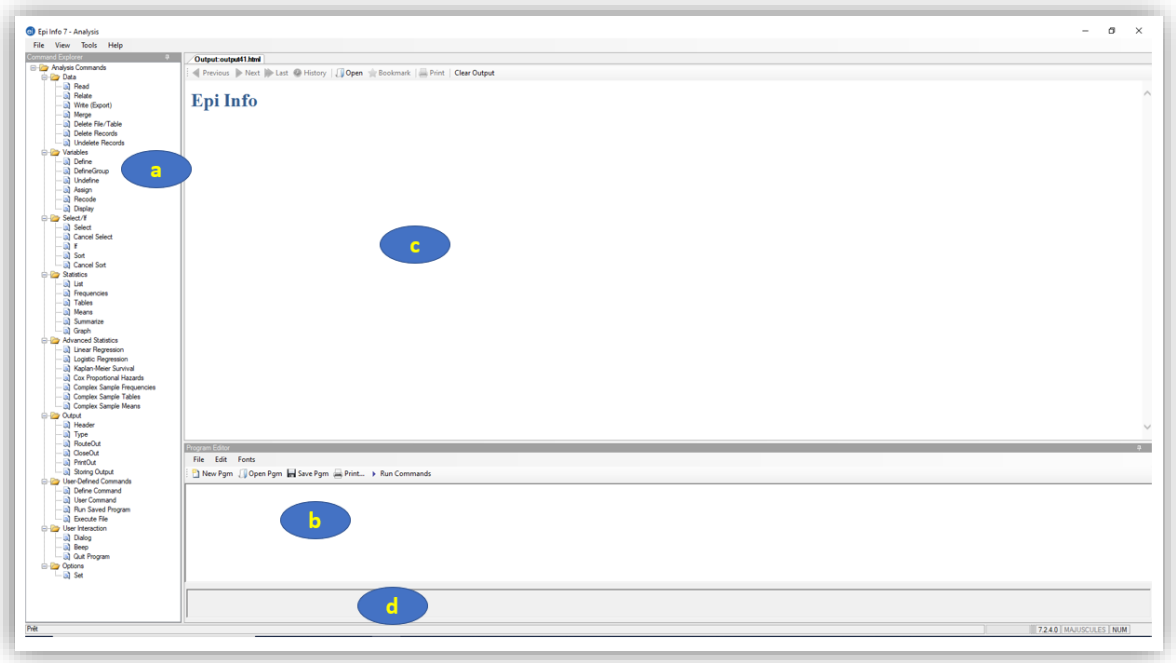


Figure 3

La fenêtre qui apparait est formée de quatre parties : la liste des commandes (Figure 3.a), l'éditeur de programme (Figure 3.b), c'est-à-dire le champ où s'écrit le programme après avoir cliqué sur une des commandes de la liste des commandes (c'est aussi dans ce champ que vous pouvez écrire vous-même votre programme une fois que vous serez à l'aise avec le code de programmation d'Epi Info !), la sortie des résultats de l'analyse statistique (Figure 3.c), la sortie pour les éventuels messages d'erreurs (Figure 3.d).

Les commandes listées dans la liste des commandes se regroupent en neuf sections (Figure 3.a). Je vais décrire très succinctement les cinq premières sections, car ce sont certaines commandes de ces sections que je vais traiter dans ce guide. Les commandes sous « Data » permettent de créer une base de données, soit à partir d'un import, soit à partir de bases de données existantes, soit à partir d'un export (Figure 4.a). Les commandes sous « Variables » permettent de créer des variables dans le fichier de données en cours de lecture (Figure 4.b). Les commandes sous « Select/if » permettent (entre autres) de travailler sur une sélection de lignes du fichier de données (Figure 4.c). Les commandes sous « Statistics » permettent de réaliser des statistiques simples (pourcentages, moyennes, médianes, tests statistiques classiques, graphiques) (Figure 4.d). Les commandes sous « Advanced statistics » permettent de réaliser des statistiques avancées (entre autres des modèles de régression et des courbes de survie de Kaplan-Meier) (Figure 4.e).

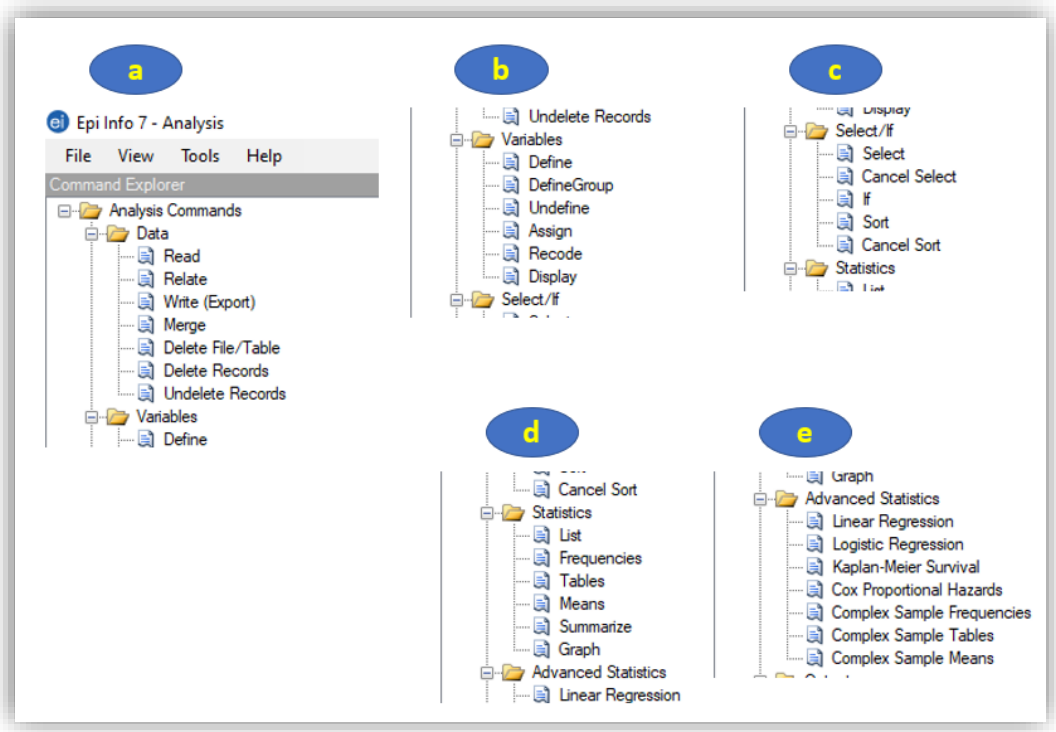


Figure 4

II. Quelques rappels indispensables

A. Ressources utilisées dans ce guide

Je ferai de temps en temps référence au cours de ce guide à quatre documents téléchargeables sur Internet : un polycopié de bases en biostatistique et un document « Utilisation d'Excel et du site Internet BiostaTGV pour réaliser quelques statistiques de base » (tous deux disponibles ici, sous « Bases en Biostatistique »), un polycopié d'analyse de survie (disponible ici, sous « Analyse de survie »), et un polycopié d'épidémiologie clinique (disponible ici, sous « Epidémiologie clinique »). Ces documents contiennent les bases théoriques de ce que je vais présenter ici. En effet, à part pour les modèles de régression, je ne présenterai ici aucun aspect théorique (ou alors, très peu), et je vous recommande de vous référer à ces documents pour ces aspects théoriques.

B. Rappel sur la structure d'un fichier de données pour analyses statistiques

Avant une analyse statistique, un fichier de données doit être structuré de façon rigoureuse pour ensuite conduire des analyses statistiques sur ces données. Pour vérifier cette structure, je vous propose de lire le document « Comment structurer un fichier de données Excel avant analyses statistiques » ici, dans la partie intitulée « Collecte des données d'une enquête épidémiologique et structure d'un fichier de données ». Vous y verrez notamment que le fichier de données doit comporter le nom des variables sur la première ligne, et les individus sont présentés en ligne.

Pour être analysable statistiquement, une variable doit être numérique. C'est-à-dire qu'elle doit être renseignée pour chaque individu sous forme d'un nombre. Ce nombre affecté à chaque individu varie selon le type de variable.

C. Définitions de critère de jugement, d'exposition, et de variable

Le « critère de jugement »[2] (abrégé « CdJ » à partir de maintenant) est l'état de santé que l'on étudie seul (de façon descriptive), ou bien dont on étudie l'association avec une ou plusieurs « expositions ».

Une « exposition » est une caractéristique intrinsèque d'un animal (âge, sexe, race, concentration en urée, …), ou extrinsèque (environnement, traitements reçus, …), qui ne soit pas le CdJ étudié.

Une « variable » représente une caractéristique d'un individu dans le fichier de donnée. Par exemple, si l'on veut savoir si, parmi les chiens, la présence d'une hypercholestérolémie est associée à la présence d'un décès dans les 3 ans suivant une consultation vétérinaire, l'exposition est la présence (*versus* absence) d'une hypercholestérolémie, le CdJ est la présence (*versus* absence) d'un décès dans les 3 ans. Dans le fichier de données, l'exposition sera représentée, par exemple, par la variable HYPERCHOLESTEROLEMIE, le CdJ sera représenté, par exemple, par la variable DECES_3ANS. Dans ce guide, j'écrirai en majuscules sans guillemets le nom des variables utilisées. Il existe quatre types de variables que je vais présenter ci-dessous. Dans toute la suite de ce guide, je ne vais quasiment plus parler que de « variable » (sauf exception), même si parfois, le terme « exposition » aurait été plus pertinent.

D. Rappel sur les différents types de variables

1. Variables binaires

Une variable binaire est une variable à deux modalités (ou classes). On peut citer comme exemple le sexe d'un animal (mâle ou femelle) ou la présence ou absence d'une maladie. Par convention, il est recommandé de coder en 0/1 dans le fichier de données une variable binaire (et en l'occurrence, pour interpréter les sorties d'Epi Info, je vous recommande très fortement de coder les variables binaires en 0/1). Le nom d'une variable binaire devrait être le nom de la modalité pour laquelle « 1 » a été attribué (c'est un conseil que je vous donne pour grandement faciliter l'interprétation des résultats fournis par le logiciel). Par exemple, si pour le sexe de l'animal, dans le fichier de données, « 1 » a été attribué aux femelles et « 0 » aux mâles, la variable devrait être nommée « Femelle » dans le fichier de données. Dans le cas d'une exposition binaire, la catégorie « exposée » est celle pour laquelle « 1 » a été attribuée à la variable correspondante, et la catégorie « non exposée » est celle pour laquelle « 0 » a été attribuée à cette même variable. De même, dans le cas d'une maladie binaire, la catégorie « malade » est celle pour laquelle « 1 » a été attribuée à la variable correspondante, et la catégorie « non malade » est celle pour laquelle « 0 » a été attribuée à cette même variable.

2. Variables qualitative nominales

Une variable qualitative nominale est une variable avec trois modalités ou plus, chacune des modalités n'étant *a priori* pas ordonnée les unes par rapport aux autres. On peut citer comme exemple la race d'un animal. Le codage d'une variable qualitative nominale peut être en 0/1/2/etc. ou bien en 1/2/3/etc.

[2] « outcome » en anglais

3. *Variables qualitative ordinales*

Une variable qualitative ordinale est une variable avec trois modalités ou plus, chacune des modalités étant ordonnée les unes par rapport aux autres. On peut citer comme premier exemple le format de la race d'un chien (petite race, race moyenne, grande race, race géante). Une variable qualitative ordinale peut aussi être obtenue à partir d'un recodage en classes d'une variable initialement quantitative, comme par exemple le temps passé par semaine à jouer avec son chien (0-30 min, 30 min – 2 heures, 2-3 heures, > 3 heures).

4. *Variables quantitatives*

Une variable quantitative est une variable continue avec un chiffre après la virgule (existant, ou possible si l'instrument de mesure était idéalement très précis) ou bien une variable discrète représentant un dénombrement. On peut citer comme exemple l'âge d'un animal, un score clinique recueilli à partir d'une échelle sur 10 points, ou bien le nombre de nodules cutanés.

III. Présentation du fichier de données fictif

Le fichier de données qui va être utilisé dans ce guide est un fichier de données fictif, provenant d'une étude prospective, fictive elle aussi, mais s'inspirant d'une précédente étude (Hua et al., 2016). Cette étude a recruté 99 chiens adultes qu'elle a suivis au cours des années à partir d'une consultation chez un vétérinaire (J0), avec un suivi minimum de 3 ans. Il n'y a eu aucun perdu de vue. Les variables contenues dans ce fichier de données sont listées ci-dessous.

DECES : variable binaire, codée en 0/1. Elle vaut « 0 » si le chien était toujours en vie à la fin de l'étude, « 1 » s'il était décédé (quelle que soit la cause du décès) au cours de l'étude.

DECES_3_ANS : variable binaire, codée en 0/1. Elle vaut « 0 » si le chien était toujours en vie 3 ans après l'inclusion dans l'étude, « 1 » s'il était décédé dans les 3 ans après l'inclusion.

SURVIE : variable quantitative représentant le délai, en années (avec 1 chiffre après la virgule), entre la date d'inclusion dans l'étude (le J0) et soit la date de fin d'étude pour les chiens toujours en vie à la fin de l'étude soit la date de décès pour les chiens décédés.

RACE_4CL : variable qualitative nominale, codée en 0/1/2/3. Elle vaut « 0 » pour les chiens de race Golden, « 1 » pour la race Labrador, « 2 » pour la race croisée Golden/Labrador, « 3 » pour une autre race.

LABRADOR : variable binaire, codée en 0/1. Elle vaut « 0 » si le chien n'était pas de race Labrador, « 1 » s'il l'était.

GOLDEN : variable binaire, codée en 0/1. Elle vaut « 0 » si le chien n'était pas de race Golden, « 1 » s'il l'était.

CROISEE : variable binaire, codée en 0/1. Elle vaut « 0 » si le chien n'était pas de race croisée Golden/Labrador, « 1 » s'il l'était.

AUTRE_RACE : variable binaire, codée en 0/1. Elle vaut « 0 » si le chien était de race Labrador, Golden, ou de race croisée Golden/Labrador, et « 1 » s'il était d'une autre race.

FEMELLE : variable binaire, codée en 0/1, « 0 » si le chien était un mâle, « 1 » s'il était une femelle.

AGE : variable quantitative représentant l'âge du chien à la consultation, en années entières.

AGE_4CL : variable qualitative ordinale, codée en 0/1/2/3 à partir des quartiles de la variable AGE. Cette variable vaut « 0 » pour les chiens avec dont l'âge est < 7 ans, « 1 » pour les chiens avec un âge compris entre 7 et 9 ans (exclu), « 2 » pour les chiens avec un âge compris entre 9 et 11 ans (exclu), et « 3 » pour les chiens avec un âge supérieur ou égal à 11 ans.

ALAT : variable quantitative représentant la concentration en ALAT, en UI/L.

UREE : variable quantitative représentant la concentration en urée, en g/L.

UREE_4CL : variable qualitative ordinale, codée en 0/1/2/3 à partir des quartiles de la variable UREE. Cette variable vaut « 0 » pour les chiens avec une concentration en urée < 0,24 g/L, « 1 » pour les chiens avec une concentration en urée comprise entre 0,24 g/L et 0,28 g/L (exclu), « 2 » pour les chiens avec une concentration en urée comprise entre 0,28 g/L et 0,33 g/L (exclu), et « 3 » pour les chiens avec une concentration en urée supérieure ou égale à 0,33 g/L.

CHOLES_3CL : variable qualitative ordinale, codée en 0/1/2. Elle vaut « 0 » pour les chiens avec une hypocholestérolémie, « 1 » pour les chiens avec une normocholestérolémie, et « 2 » pour les chiens avec une hypercholestérolémie.

HYPER_CHOLES : variable binaire, codée en 0/1. Elle vaut « 0 » si le chien ne présentait pas d'hypercholestérolémie, et « 1 » s'il en présentait une.

DEMARCHE_ANORMALE : variable binaire, codée en 0/1, « 0 » si le chien avait une démarche normale, « 1 » s'il avait une démarche anormale.

CONSTANTE : variable créée pour l'utilisation d'EI, valant « 1 » pour tous les chiens de l'échantillon (utile uniquement dans le cas de la réalisation d'un courbe de Kaplan-Meier globale).

IV. Importer les données à analyser et sauvegarder un programme

A. Création préalable d'un « projet »

Avant de commencer à importer un fichier Excel pour ensuite l'analyser statistiquement dans Epi Info, je vous recommande de créer un « projet » au préalable. Cela vous permettra d'enregistrer votre programme comprenant toutes vos analyses statistiques. Pour cela, après avoir lancé Epi Info, on clique sur « Create Forms » (cf. Figure 5.a).

Figure 5

Ensuite, on clique sur « New project » (Figure 6.a), on remplit le champ « Name » avec un nom de votre choix (Figure 6.b), on clique sur « Browse » (Figure 6.c) pour choisir d'enregistrer ce projet à l'emplacement de votre choix, on met ce que l'on veut dans le champ « Form name » (Figure 6.d), puis on clique sur « Ok » (Figure 6.e).

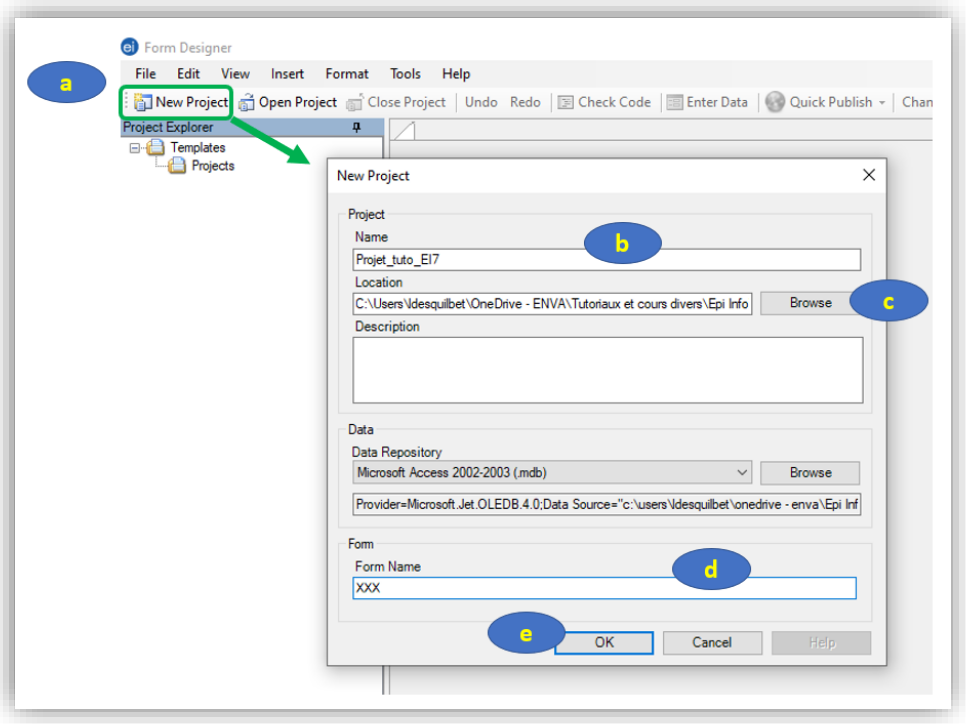


Figure 6

On ferme ensuite la fenêtre qui apparait, en cliquant sur « File » → « Exit » (cf. Figure 7).

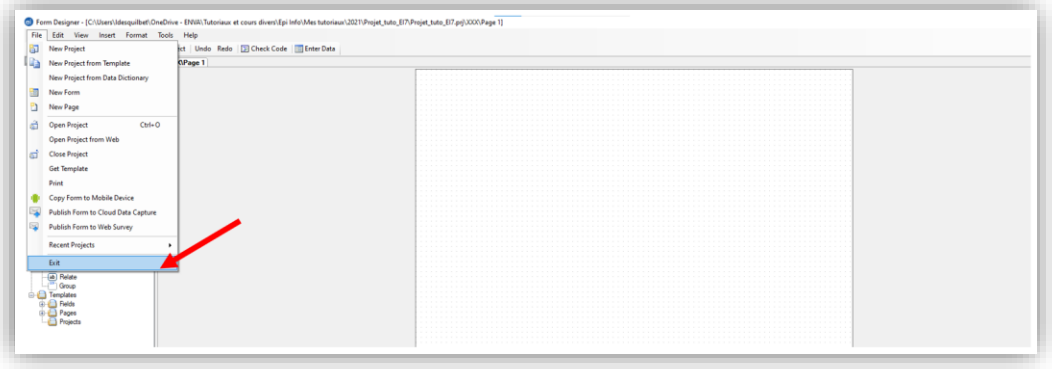


Figure 7

B. Importation de données d'un fichier Excel (ou sous un autre format)

Avant d'importer un fichier de données, vous devez vous assurer de l'avoir enregistré sous format .xls (ce qui est possible si vous utilisez le logiciel LibreOffice Calc). Pour importer un fichier .xls, tout comme pour réaliser les analyses statistiques, on clique sur « Classic » dans la fenêtre de démarrage d'Epi Info (cf. Figure 8).

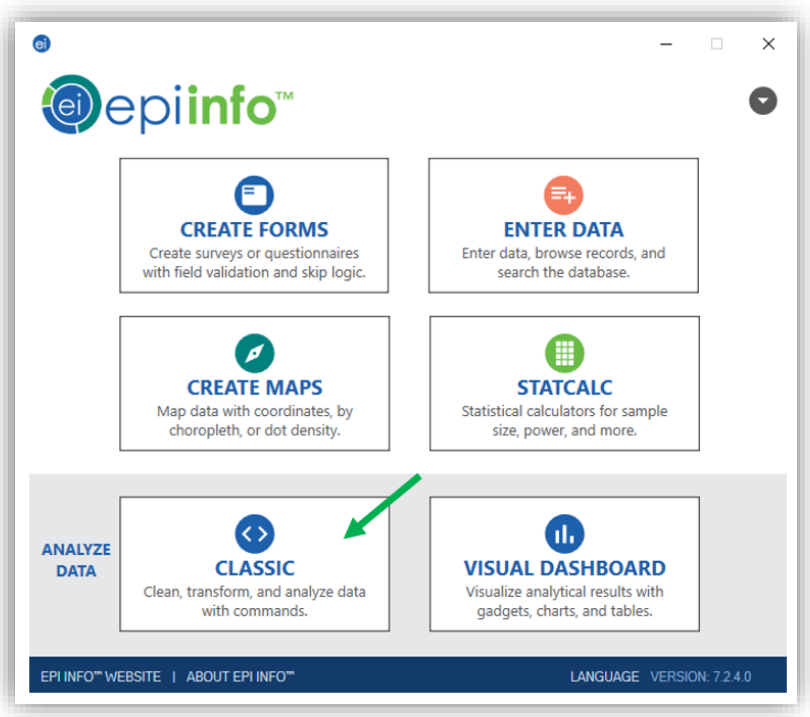


Figure 8

On clique ensuite sur « Read » (Figure 9.a), on sélectionne « Microsoft Excel 97-2003 » (Figure 9.b), on va chercher le fichier de données .xls en cliquant sur « Browse » deux fois ((Figure 9.c) et (Figure 9.d)), on clique sur « First row contains header information » (parce que la première ligne du fichier de données comporte le nom des variables ; Figure 9.e), on clique sur « Ok » (Figure 9.f), on sélectionne l'onglet qui comprend les données que l'on souhaite analyser (Figure 9.g), et on clique sur « Ok » (Figure 9.h).

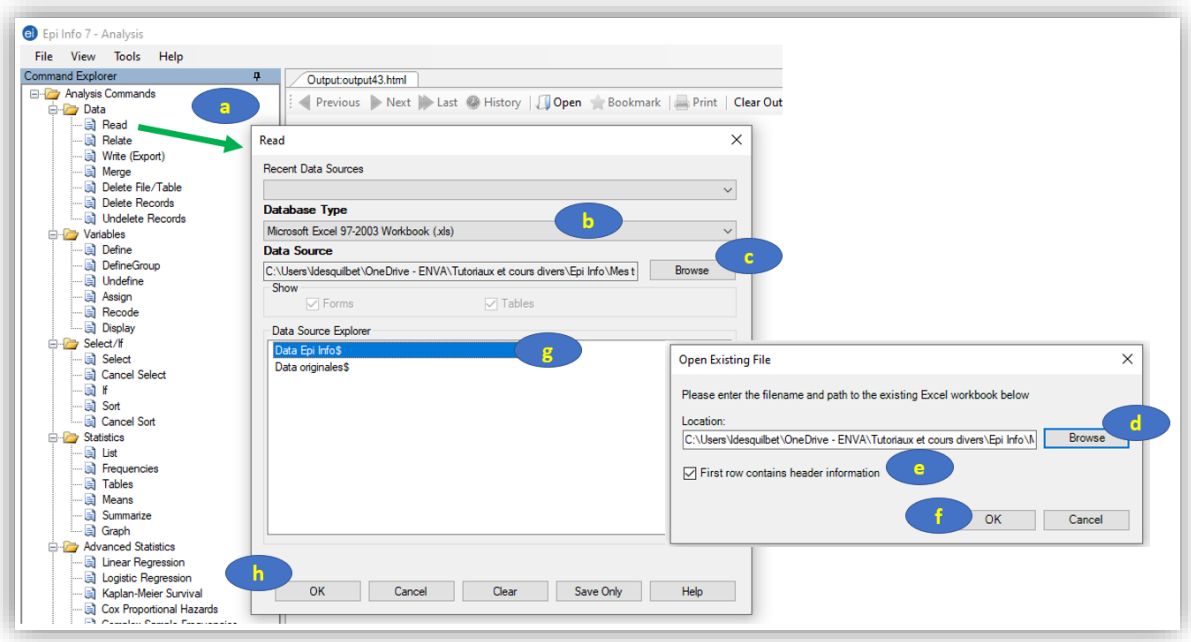


Figure 9

Dans la sortie des résultats (Figure 10.a), Epi Info indique que votre fichier a été importé, et qu'il comprend 99 individus (devant « Record Count »). Le programme d'import de vos données a été automatiquement écrit dans la fenêtre d'éditeur de programme (Figure 10.b).

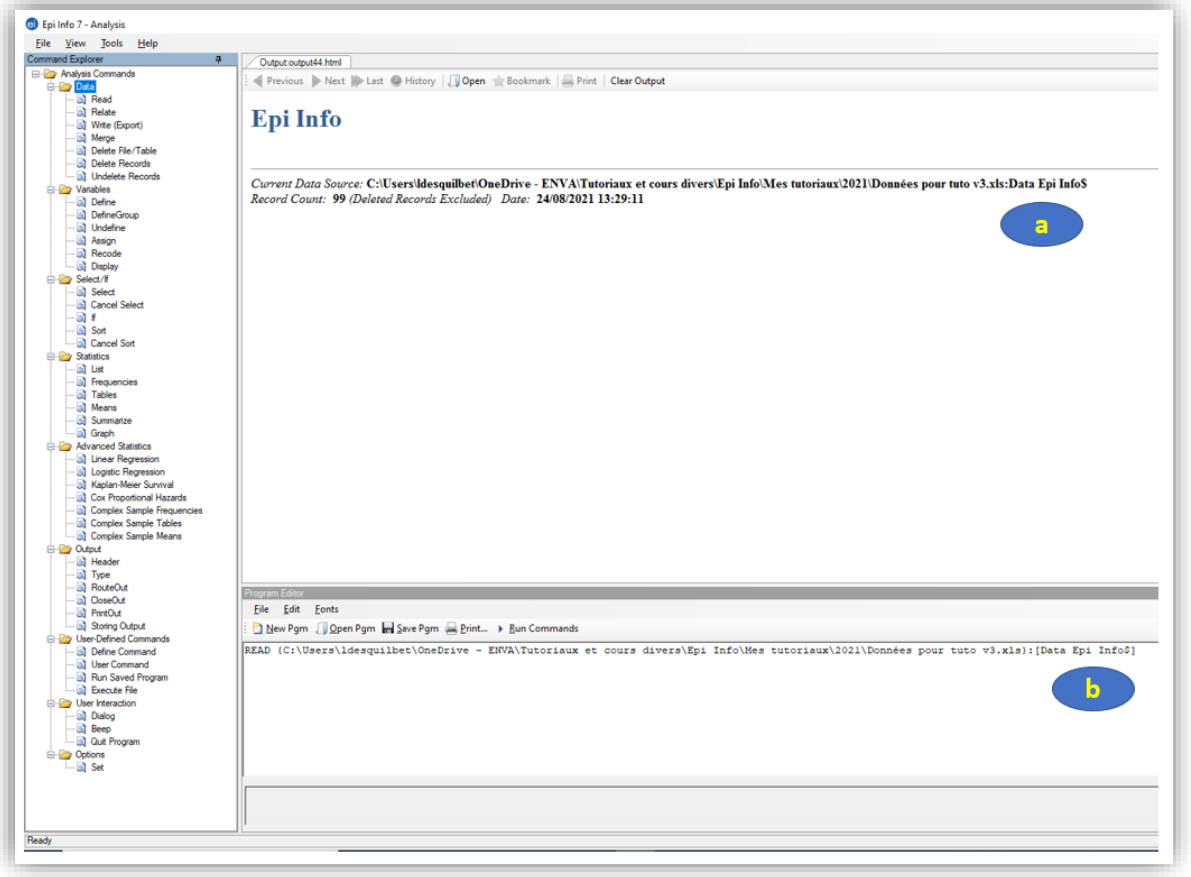


Figure 10

C. Sauvegarder son programme

Pour éviter de refaire la même (et longue) manipulation d'import de données, pour éviter de retaper systématiquement votre programme, et/ou pour garder une trace (indispensable) de toutes vos analyses que vous avez réalisées sur un fichier de données, il est indispensable d'enregistrer votre programme. Pour cela, on clique sur « Save pgm » dans l'éditeur de programme (Figure 11.a), on sélectionne son projet (Figure 11.b), on tape le nom du programme de son choix (Figure 11.c), puis on clique sur « Ok » (Figure 11.d).

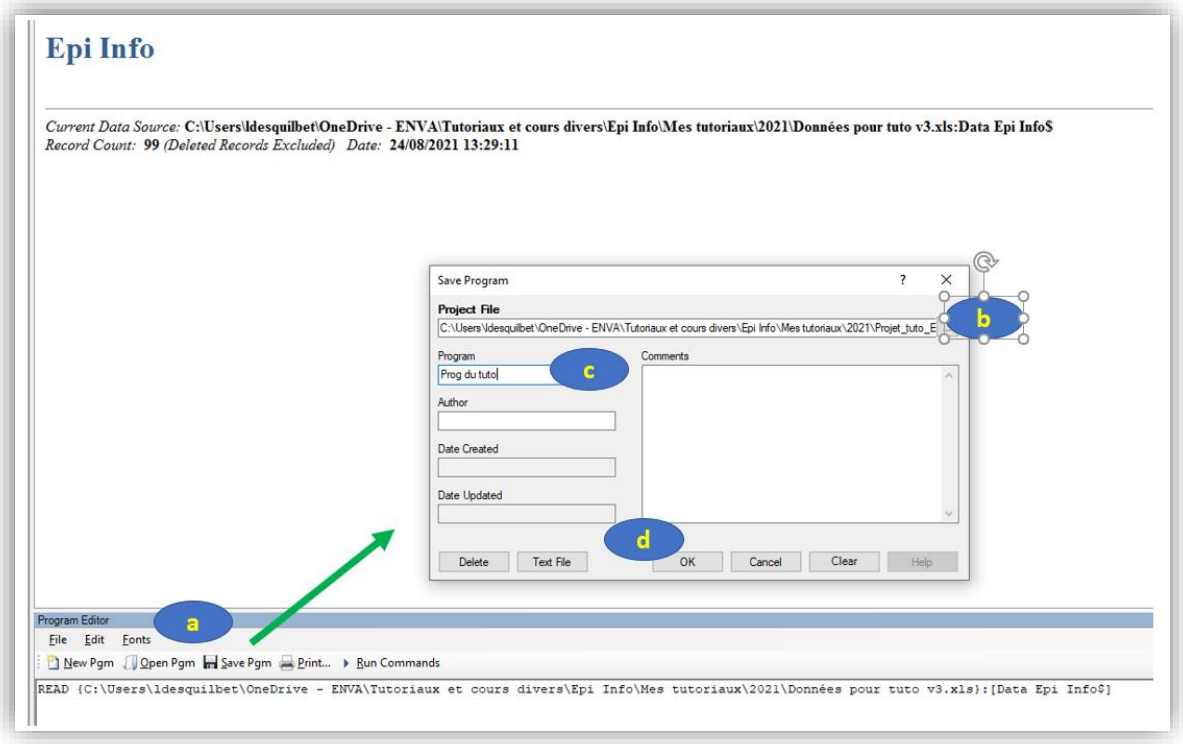


Figure 11

Une fois cette première étape d'enregistrement réalisée, on n'aura qu'à cliquer sur « Save pgm » au fur et à mesure des analyses (Figure 12.a) (vous devrez néanmoins suivre à nouveau les étapes décrites sur la Figure 11).

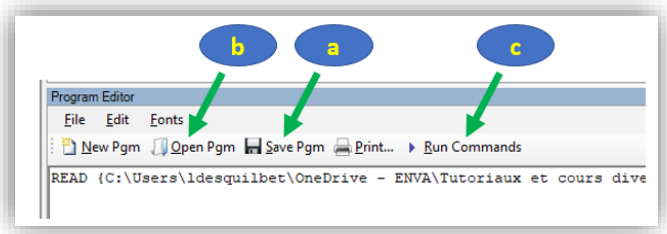


Figure 12

A la fin de la journée de travail, on enregistre le programme une dernière fois. Le lendemain, pour reprendre le travail, on cliquera sur « Open Pgm » (Figure 12.b) (et on sélectionnera le projet puis le programme enregistré la veille). Ensuite, pour exécuter certaines lignes du programme (et notamment, la première qui correspond à l'import du fichier de données au moment de recommencer à utiliser Epi Info après une bonne nuit de sommeil), on sélectionne la ligne de commande à exécuter (en double-cliquant dessus), puis on clique sur « Run Commands » (Figure 12.c).

Chapitre 2 – Analyses statistiques de base

I. Introduction

A. Vérifications préliminaires avant les analyses statistiques

Avant de commencer à faire des analyses statistiques, il est indispensable de vérifier que les variables sur lesquelles vous comptez travailler sont bien des variables numériques. Pour cela, on clique sur « Display » (Figure 13.a), puis sur « Ok » (Figure 13.b).

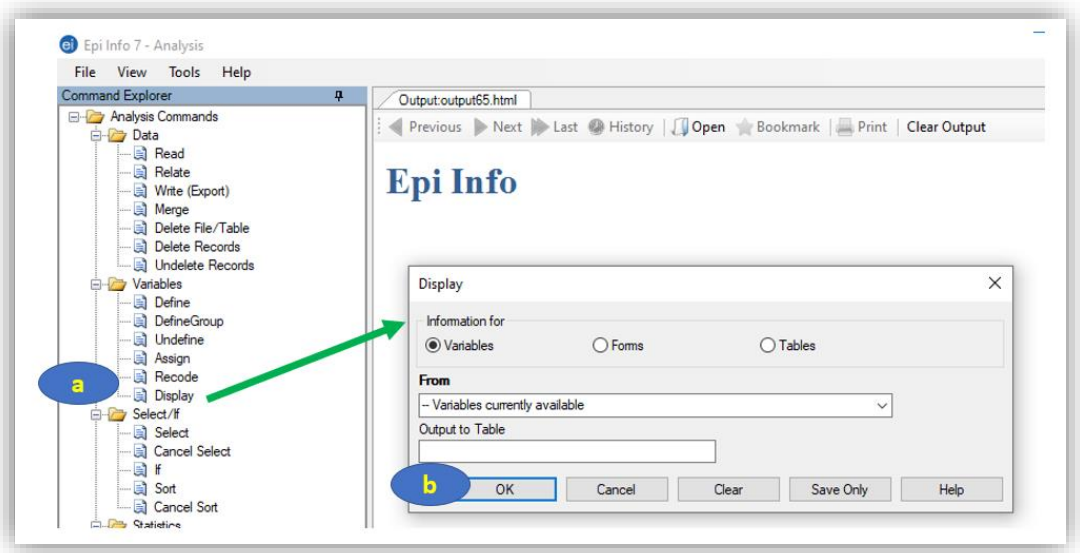
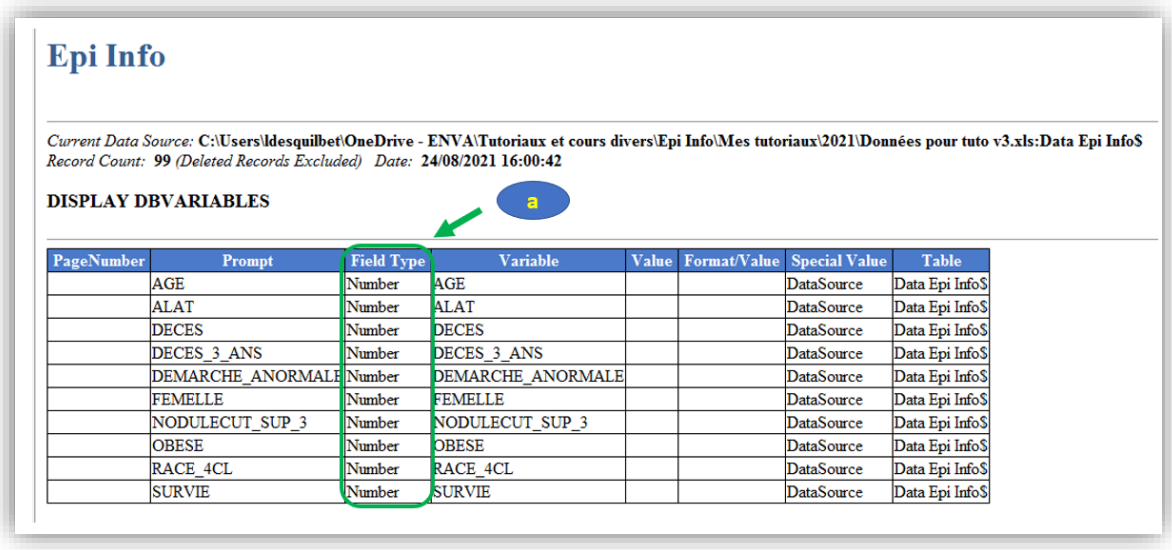


Figure 13

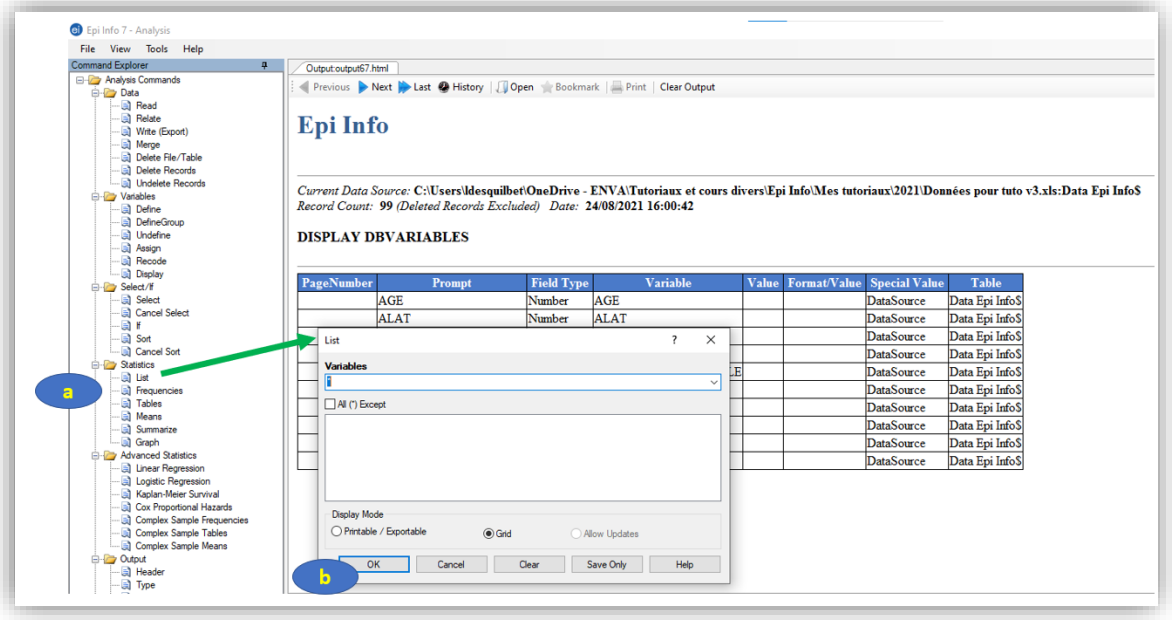
On obtient le résultat ci-dessous. Dans la colonne « Field Type » (Figure 14.a), on voit bien que toutes les variables du fichier de données sont bien numériques (« Number » indiqué dans la colonne). Il est fondamental de bien vérifier que toutes les variables que vous souhaitez utiliser dans vos analyses statistiques sont effectivement numériques. Si tel n'est pas le cas pour une variable, alors vous devez retourner sur votre fichier Excel de données, et trouver la raison pour laquelle Epi Info considère la variable comme du texte, et non pas de façon numérique. Sachez que nettoyer un fichier de données pour que toutes les variables que vous voulez utiliser pour vos analyses statistiques soient numériques est une tâche qui prend souvent beaucoup de temps, et une tâche sur laquelle on avait tout sauf prévu de passer du temps !



PageNumber	Prompt	Field Type	Variable	Value	Format/Value	Special Value	Table
	AGE	Number	AGE			DataSource	Data Epi Info$
	ALAT	Number	ALAT			DataSource	Data Epi Info$
	DECES	Number	DECES			DataSource	Data Epi Info$
	DECES_3_ANS	Number	DECES_3_ANS			DataSource	Data Epi Info$
	DEMARCHE_ANORMALE	Number	DEMARCHE_ANORMALE			DataSource	Data Epi Info$
	FEMELLE	Number	FEMELLE			DataSource	Data Epi Info$
	NODULECUT_SUP_3	Number	NODULECUT_SUP_3			DataSource	Data Epi Info$
	OBESE	Number	OBESE			DataSource	Data Epi Info$
	RACE_4CL	Number	RACE_4CL			DataSource	Data Epi Info$
	SURVIE	Number	SURVIE			DataSource	Data Epi Info$

Figure 14

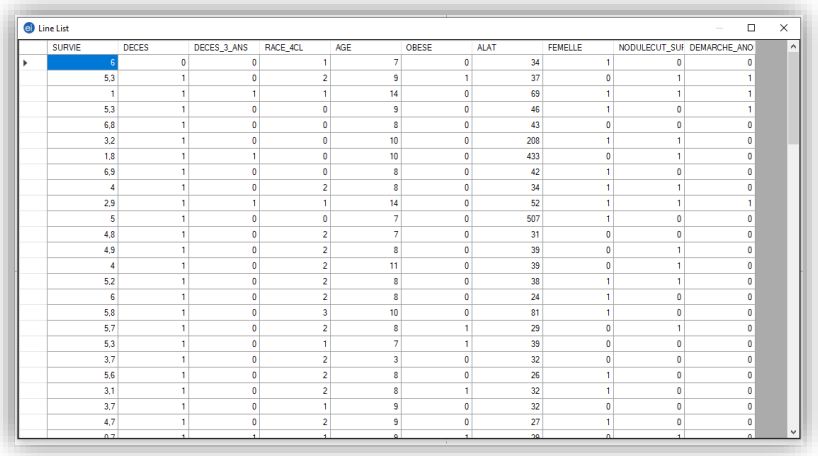
Ensuite, on pourrait vouloir vérifier visuellement la structure et le contenu du fichier de données qui a été importé. Pour cela, on clique sur « List » (Figure 15.a), puis sur « Ok » (Figure 15.b).



PageNumber	Prompt	Field Type	Variable	Value	Format/Value	Special Value	Table
	AGE	Number	AGE			DataSource	Data Epi Info$
	ALAT	Number	ALAT			DataSource	Data Epi Info$
						DataSource	Data Epi Info$
						DataSource	Data Epi Info$
						DataSource	Data Epi Info$
						DataSource	Data Epi Info$
						DataSource	Data Epi Info$
						DataSource	Data Epi Info$
						DataSource	Data Epi Info$
						DataSource	Data Epi Info$

Figure 15

On obtient la base de données sur laquelle on va travailler (cf. Figure 16).



	SURVIE	DECES	DECES_3_ANS	RACE_4CL	AGE	OBESE	ALAT	FEMELLE	NODULECUT_SUF	DEMARCHE_ANO
▸	6	0	0	1	7	0	34	1	0	0
	5.3	1	0	2	9	1	37	0	1	1
	1	1	1	1	14	0	69	1	1	1
	5.3	1	0	0	9	0	46	1	0	1
	6.8	1	0	0	8	0	43	0	0	0
	3.2	1	0	0	10	0	208	1	1	0
	1.8	1	1	0	10	0	433	0	1	0
	6.9	1	0	0	8	0	42	1	0	0
	4	1	0	2	8	0	34	1	1	0
	2.9	1	1	1	14	0	52	1	1	1
	5	1	0	0	7	0	507	1	0	0
	4.8	1	0	2	7	0	31	0	0	0
	4.9	1	0	2	8	0	39	0	1	0
	4	1	0	2	11	0	39	0	1	0
	5.2	1	0	2	8	0	38	1	1	0
	6	1	0	2	8	0	24	1	0	0
	5.8	1	0	3	10	0	81	1	0	0
	5.7	1	0	2	8	1	29	0	1	0
	5.3	1	0	1	7	1	39	0	0	0
	3.7	1	0	2	3	0	32	0	0	0
	5.6	1	0	2	8	0	26	1	0	0
	3.1	1	0	2	8	1	32	1	0	0
	3.7	1	0	1	9	0	32	0	0	0
	4.7	1	0	2	9	0	27	1	0	0

Figure 16

Maintenant, nous sommes prêts à réaliser les analyses statistiques à partir des données, qui sont « propres », du fichier de données.

B. Convention de présentation des résultats issus des copies d'écran d'Epi Info

Lorsque je vais écrire les résultats statistiques qu'Epi Info fournit, je ne vais volontairement faire aucun arrondi, pour que vous puissiez repérer tout de suite où se trouve, dans la copie d'écran, les chiffres que je mentionne. Une exception néanmoins à cela : lorsque je mentionnerai les Odds Ratios ou Hazard Ratio fournis par le logiciel dans le Chapitre III sur les modèles de régression, que j'arrondirai à deux chiffres après la virgule.

II. Statistiques descriptives

A. Décrire une variable binaire ou qualitative

Pour décrire une variable binaire ou qualitative, nous allons utiliser le « tableau de fréquences », qui est un tableau qui présente les effectifs et les pourcentages pour chaque modalité d'une variable binaire ou qualitative. Nous allons prendre pour l'exemple la variable RACE_4CL. Pour obtenir le tableau de fréquences, on clique sur « Frequencies » (Figure 17.a), on sélectionne la variable RACE_4CL dans la liste déroulante (Figure 17.b), puis on clique sur « Ok » (Figure 17.c).

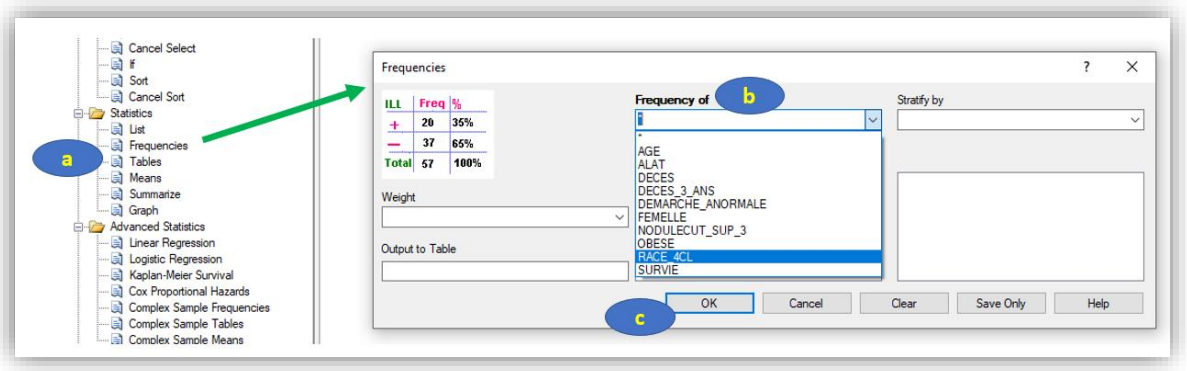
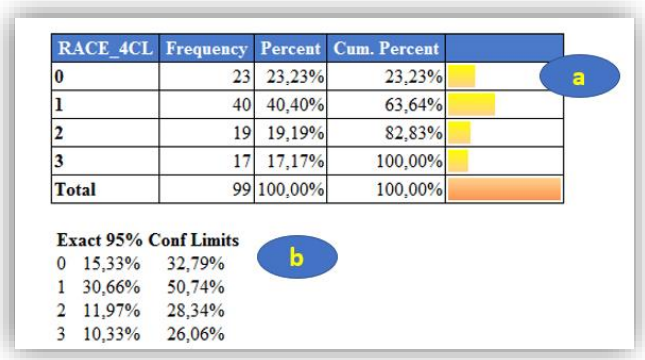


Figure 17

Le résultat est présenté sur la Figure 18. On peut lire que les chiens de race Golden (RACE_4CL = 0) sont au nombre de 23, ce qui représente 23,23% de l'échantillon des 99 chiens (Figure 18.a). L'intervalle de confiance à 95% (noté « $IC_{95\%}$ » dans toute la suite de ce guide) de chaque pourcentage est fourni par Epi Info (Figure 18.b). Celui du pourcentage de Golden de 23,23% est : $[15{,}33\% ; 32{,}79\%]_{95\%}$.



RACE_4CL	Frequency	Percent	Cum. Percent
0	23	23,23%	23,23%
1	40	40,40%	63,64%
2	19	19,19%	82,83%
3	17	17,17%	100,00%
Total	99	100,00%	100,00%

a

Exact 95% Conf Limits

0	15,33%	32,79%
1	30,66%	50,74%
2	11,97%	28,34%
3	10,33%	26,06%

b

Figure 18

B. Décrire une variable quantitative

Pour vous montrer comment obtenir avec Epi Info les différents indicateurs statistiques décrivant une variable quantitative, nous allons prendre pour l'exemple la variable ALAT. Pour cela, on clique sur « Means » (Figure 19.a), on sélectionne la variable « ALAT » (Figure 19.b), puis on clique sur « Ok » (Figure 19.c).

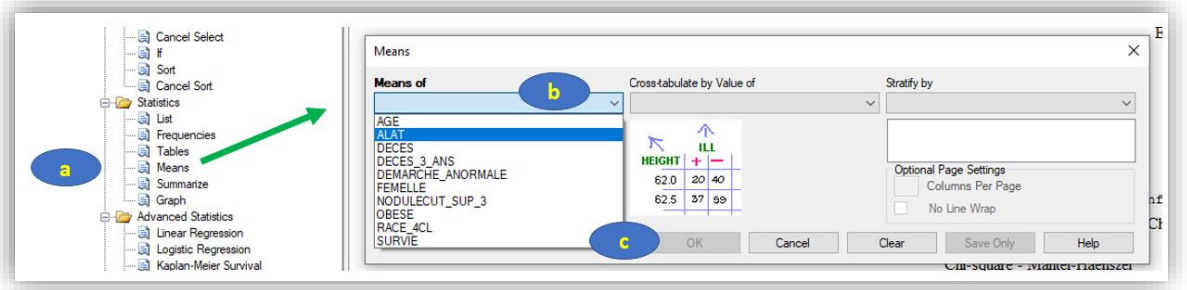


Figure 19

On obtient alors le résultat présenté sur la Figure 20. On observe ainsi qu'il y a 98 données non manquantes sur les ALAT (« Obs = 98,000 »), donc il y a une donnée manquante sur ce paramètre biologique (puisqu'il y a 99 chiens dans l'échantillon). La moyenne de la concentration en ALAT est de 57,6398 UI/L, la Standard Deviation[3] (SD) de 69,6023 UI/L, le minimum de 11 UI/L, le 1er quartile (25ème percentile) de 32 UI/L, la médiane de 39 UI/L, le 3ème quartile (75ème percentile) de 53 UI/L, et le maximum de 507 UI/L.

MEANS ALAT

Obs	Total	Mean	Variance	Std Dev
98,0000	5648,7000	57,6398	4844,4822	69,6023

Minimum	25%	Median	75%	Maximum	Mode
11,0000	32,0000	39,0000	53,0000	507,0000	32,0000

Figure 20

Je n'ai pas trouvé le moyen de dresser un histogramme avec Epi Info. C'est dommage. Pour cela, je vous recommande la lecture de la partie « Vérifier la normalité d'une variable quantitative » dans le document « Utilisation d'Excel et du site Internet BiostaTGV ».

III. Association statistique entre deux variables

A. Croisement de deux variables binaires ou qualitatives, et test statistique

1. *Introduction*

Le croisement de deux variables (binaires ou qualitatives) permet d'étudier l'association entre ces deux variables. Je ne reviendrai pas dans ce guide sur la façon de correctement lire un tableau croisant deux variables (notamment, savoir faire la distinction entre les « bons » et les « mauvais » pourcentages à citer). Si ce n'est pas clair pour vous, relisez certaines parties du polycopié de biostatistique.

2. *Croisement de deux variables binaires*

Nous allons prendre pour l'exemple les deux variables suivantes : FEMELLE et DECES_3_ANS. Pour croiser ces deux variables, on clique sur « Tables » (Figure 21.a), on sélectionne ensuite sous « exposure variable » la variable qui joue le rôle d'exposition (ici, FEMELLE ; Figure 21.b) et sous « outcome variable » celle qui joue le rôle du CdJ (ici, DECES_3_ANS ; Figure 21.c), puis on clique sur « Ok » (Figure 21.d).

[3] Cf. polycopié de biostatistique.

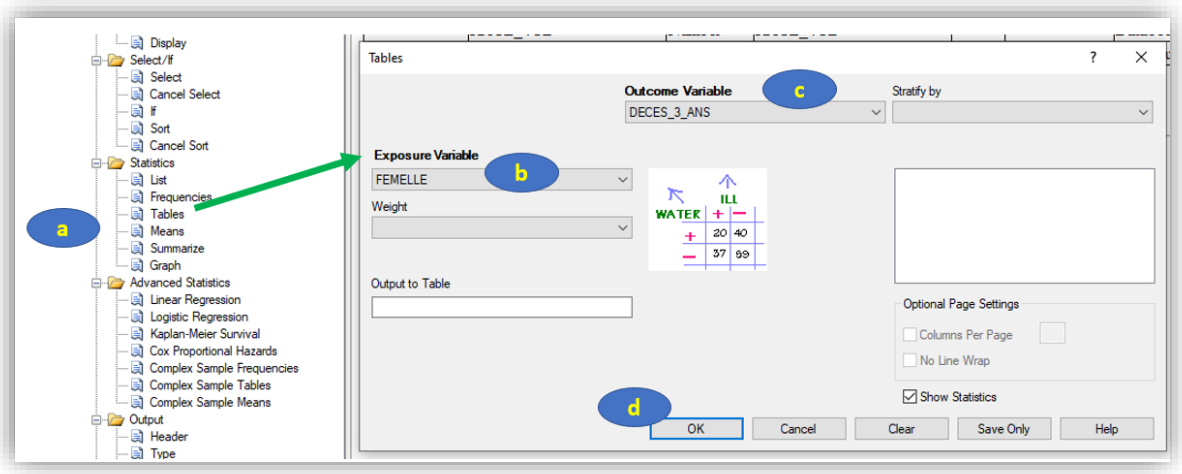
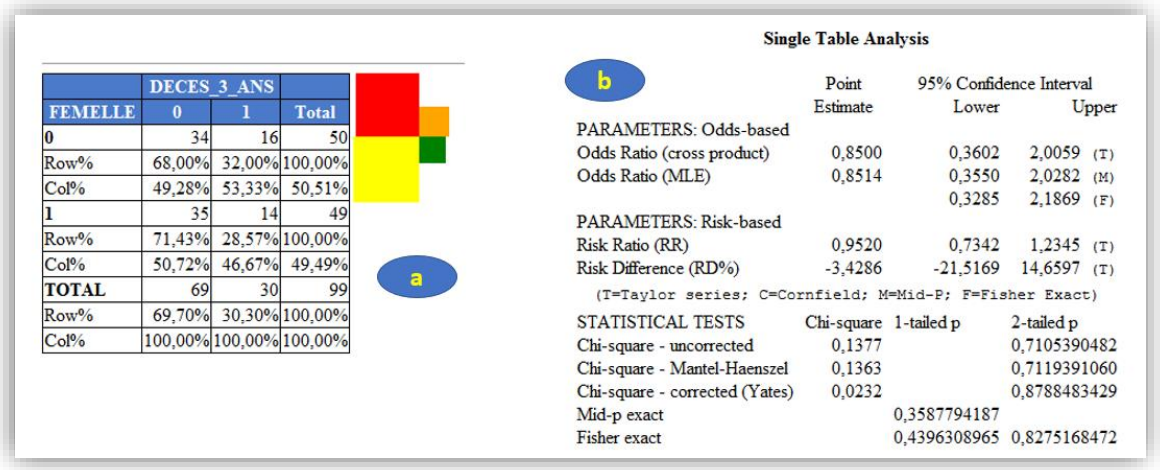


Figure 21

On obtient alors la sortie de résultats sur la Figure 22, qui comporte deux parties : le tableau d'effectifs et de pourcentages (Figure 22.a) et d'autres indicateurs statistiques (Figure 22.b).



	DECES_3_ANS		
FEMELLE	0	1	Total
0	34	16	50
Row%	68,00%	32,00%	100,00%
Col%	49,28%	53,33%	50,51%
1	35	14	49
Row%	71,43%	28,57%	100,00%
Col%	50,72%	46,67%	49,49%
TOTAL	69	30	99
Row%	69,70%	30,30%	100,00%
Col%	100,00%	100,00%	100,00%

a

b

Single Table Analysis

	Point Estimate	95% Confidence Interval Lower	Upper	
PARAMETERS: Odds-based				
Odds Ratio (cross product)	0,8500	0,3602	2,0059	(T)
Odds Ratio (MLE)	0,8514	0,3550	2,0282	(M)
		0,3285	2,1869	(F)
PARAMETERS: Risk-based				
Risk Ratio (RR)	0,9520	0,7342	1,2345	(T)
Risk Difference (RD%)	-3,4286	-21,5169	14,6597	(T)

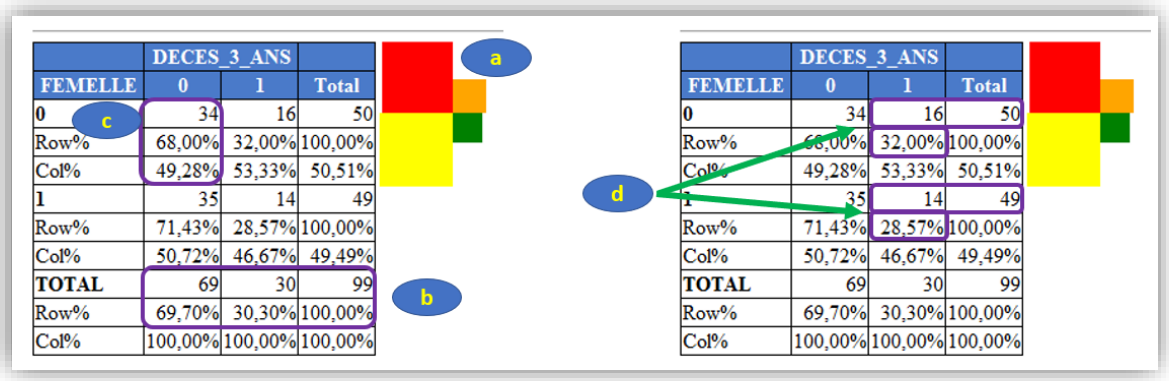
(T=Taylor series; C=Cornfield; M=Mid-P; F=Fisher Exact)

STATISTICAL TESTS	Chi-square	1-tailed p	2-tailed p
Chi-square - uncorrected	0,1377		0,7105390482
Chi-square - Mantel-Haenszel	0,1363		0,7119391060
Chi-square - corrected (Yates)	0,0232		0,8788483429
Mid-p exact		0,3587794187	
Fisher exact		0,4396308965	0,8275168472

Figure 22

Je vais tout d'abord me focaliser sur le tableau d'effectifs et de pourcentages. Tout d'abord, Epi Info fournit une représentation graphique (sommaire) de la répartition des effectifs dans chacune des cases du tableau (Figure 23.a). La surface des carrés de couleurs est proportionnelle aux effectifs dans chacune des quatre cases : 34 chiens non décédés dans les trois ans (rouge), 16 chiens décédés dans les trois ans (orange), 35 chiennes non décédées dans les trois ans (jaune), et 14 chiennes décédées dans les trois ans (vert). Les effectifs (b) de la Figure 23 nous renseignent sur le fait qu'il y a en tout 69 chiens non décédés dans les trois ans, et 30 chiens décédés dans les trois ans (avec un total bien évidemment de 99 chiens). Ensuite, l'indication « Row% » (Figure 23.c) signifie que les pourcentages sur cette ligne sont des pourcentages en ligne (donc lecture horizontale) : parmi les 50 chiens mâles (FEMELLE = 0), 34 ne sont pas décédés dans les trois ans, et 34/50 = 68,00%. Enfin, l'indication « Col% » (Figure 23.c) signifie que les pourcentages dans cette colonne sont des pourcentages en colonne (donc lecture verticale) : parmi les 69 chiens non décédés dans les trois ans, 34 sont des mâles, et 34/69 = 49,28%. Pour savoir si ces deux variables binaires sont associées, quatre couples de pourcentages peuvent être calculés et être comparés. Je vais choisir les deux pourcentages (en ligne) suivants : le pourcentage de chiens décédés parmi les chiens mâles (16/50=32,00%) et le pourcentage de chiens décédés parmi les chiens femelles (14/49=28,57%) (Figure 23.d). Dans l'échantillon, le pourcentage de chiens décédés est légèrement inférieur parmi les chiens femelles que parmi les chiens mâles (28,57%

< 32,00%). Numériquement, ces deux pourcentages étant très proches, dans l'échantillon, l'association entre le sexe du chien et le fait qu'il soit décédé dans les trois ans est quasi inexistante.



	DECES_3_ANS		
FEMELLE	0	1	Total
0	34	16	50
Row%	68,00%	32,00%	100,00%
Col%	49,28%	53,33%	50,51%
1	35	14	49
Row%	71,43%	28,57%	100,00%
Col%	50,72%	46,67%	49,49%
TOTAL	69	30	99
Row%	69,70%	30,30%	100,00%
Col%	100,00%	100,00%	100,00%

	DECES_3_ANS		
FEMELLE	0	1	Total
0	34	16	50
Row%	68,00%	32,00%	100,00%
Col%	49,28%	53,33%	50,51%
1	35	14	49
Row%	71,43%	28,57%	100,00%
Col%	50,72%	46,67%	49,49%
TOTAL	69	30	99
Row%	69,70%	30,30%	100,00%
Col%	100,00%	100,00%	100,00%

Figure 23

Passons maintenant à la sortie des autres indicateurs statistiques. La partie (a) de la Figure 24 concerne les tests statistiques testant l'association entre les deux variables (FEMELLE et DECES_3_ANS). Vous voyez que trois lignes sont dévolues au test du Chi-2, une ligne au test du mid-p, et une ligne au test exact de Fisher. Si Epi Info ne mentionne pas en bas de la sortie (a) de la Figure 24 « At least one cell has expected size <5. Chi-square may not be a valid test. » (ce qui est le cas ici), alors vous regardez le résultat de la ligne « Chi-square – uncorrected », colonne « 2-tailed p »[4]. La valeur du degré de signification testant la différence entre les deux pourcentages cités ci-dessus (32,00% et 28,57%), et provenant du test du Chi-2 classique, est de 0,7105. Ces deux pourcentages sont donc non significativement différents[5]. Si Epi Info avait mentionné « At least one cell has expected size <5. Chi-square may not be a valid test. », alors il aurait fallu lire le degré de signification à la 5ème ligne « Fisher exact » (toujours dans la même colonne), de valeur 0,8275, provenant du test statistique exact de Fisher.

[4] Ce qui signifie que les valeurs dans cette colonne sont celles du degré de signification en considérant le test statistique comme bilatéral.

[5] Je vais fixer le risque d'erreur de 1ère espèce α à 0,05 (5%) dans tout ce guide.

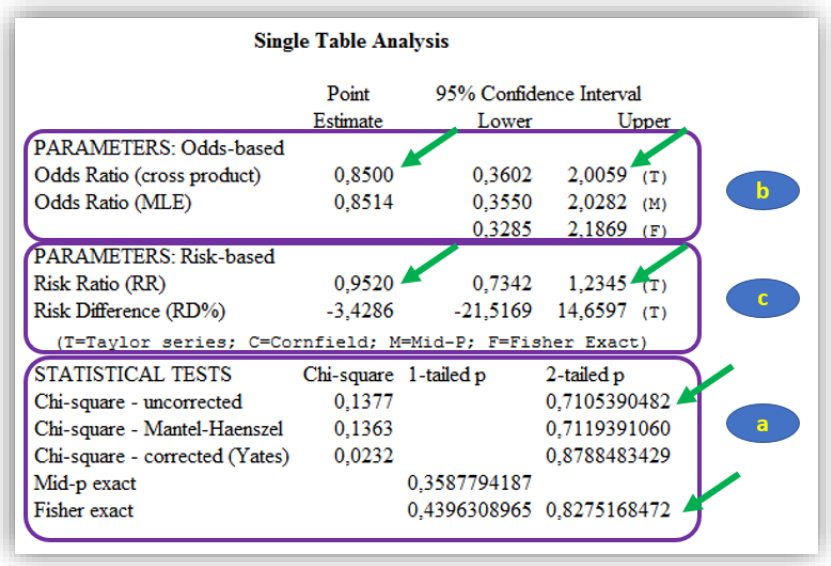


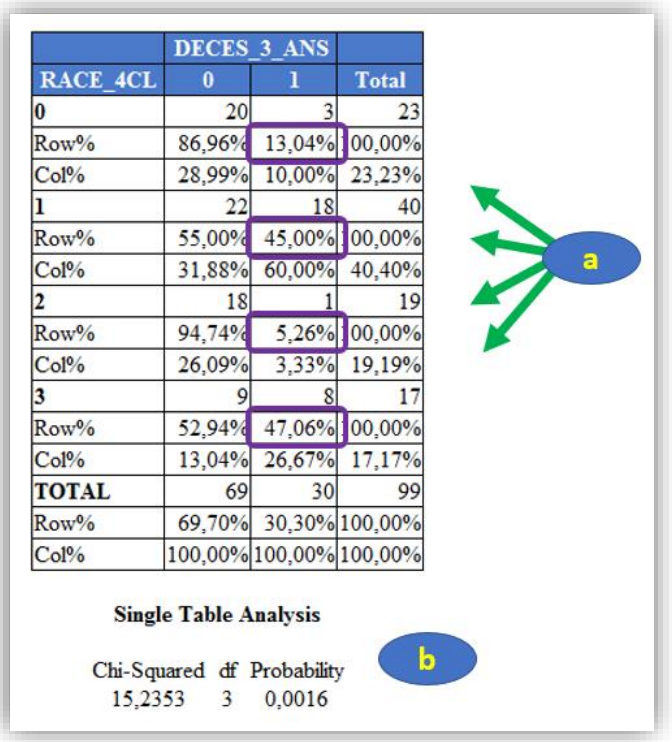
Figure 24

Epi Info fournit aussi les indicateurs statistiques qui permettent de quantifier l'association entre deux variables binaires : l'Odds Ratio (OR) et le Risque Relatif (RR). Je n'entrerai pas dans les détails du calcul de ces deux indicateurs, mais si cela vous intéresse, vous pourrez lire les parties du polycopié d'épidémiologie clinique qui sont consacrées à ces indicateurs. Dans la partie (b) de la Figure 24, il y a deux OR. En fait, c'est parce qu'Epi Info calcule l'OR de deux façons différentes : le produit en croix (« cross product »), et par l'estimation du maximum de vraisemblance (MLE[6]). Je vous recommande l'estimation selon le produit en croix. Ainsi, dans la mesure où la catégorie « exposée » est celle des chiens femelles (car le « 1 » pour la variable FEMELLE concerne les chiens femelles ; cf. Chapitre 1, Partie II.D.2, page 8), et dans la mesure où les chiens présentant le CdJ sont les chiens décédés dans les trois ans (car le « 1 » pour la variable DECES_3_ANS concerne les chiens décédés), l'$OR_{\text{Femelles } versus \text{ mâles}} = 0{,}8500\ [0{,}3602\ ;\ 2{,}0059]_{95\%}$. Cet OR est plus petit que « 1 », et c'est normal, puisque nous avions vu précédemment que la proportion de chiens présentant le CdJ (les chiens décédés) était (légèrement) *inférieure* parmi les chiens « exposés » (les chiens femelles) que parmi les chiens « non exposés » (les chiens mâles). La valeur du RR correspondante est de $0{,}9520\ [0{,}7342\ ;\ 1{,}2345]_{95\%}$ (Figure 24.c).

3. *Croisement d'une variable binaire avec une variable qualitative*

Nous allons prendre pour l'exemple les deux variables suivantes : RACE_4CL et DECES_3_ANS. Pour croiser ces deux variables, nous utilisons la même démarche que celle décrite dans la Figure 21, sauf que nous sélectionnons la variable RACE_4CL au lieu de la variable FEMELLE. Nous obtenons les résultats présentés sur la Figure 25.

[6] MLE = maximum likelihood estimation (plus d'infos ici)



	DECES_3_ANS		
RACE_4CL	0	1	Total
0	20	3	23
Row%	86,96%	13,04%	00,00%
Col%	28,99%	10,00%	23,23%
1	22	18	40
Row%	55,00%	45,00%	00,00%
Col%	31,88%	60,00%	40,40%
2	18	1	19
Row%	94,74%	5,26%	00,00%
Col%	26,09%	3,33%	19,19%
3	9	8	17
Row%	52,94%	47,06%	00,00%
Col%	13,04%	26,67%	17,17%
TOTAL	69	30	99
Row%	69,70%	30,30%	100,00%
Col%	100,00%	100,00%	100,00%

Figure 25

Nous pouvons déjà observer qu'il y a beaucoup moins de résultats statistiques que sur la Figure 22 ! Les pourcentages qui doivent être comparés pour savoir s'il existe une association entre la race et la présence d'un décès sont les pourcentages suivants : le pourcentage de chiens décédés parmi les chiens de race Golden (3/23 = 13,04%), le pourcentage de chiens décédés parmi les chiens de race Labrador (18/40 = 45,00%), le pourcentage de chiens décédés parmi les chiens de race croisée Golden/Labrador (1/19 = 5,26%), et le pourcentage de chiens décédés parmi les chiens d'autre race (8/17 = 47,06% ; Figure 25.a).

Dans la mesure où Epi Info n'indique pas « An expected value is < 5. Chi-squared may not be a valid test. », le test du Chi-2 est valide, et son degré de signification se lit sur la sortie : 0,0016 (Figure 25.b). (Si au moins un des effectifs attendus avait été inférieur à 5, alors il aurait fallu réaliser le test statistique de Fisher, qui n'est pas proposé par Epi Info dans le cas de croisement d'une variable binaire avec une variable qualitative. Dans ce cas-là, je vous recommande d'utiliser le site Internet BiostaTGV[7].) Le degré de signification étant inférieur à 0,05, on peut dire qu'il existait dans l'échantillon une association significative entre la race et le décès du chien.

4. Croisement de deux variables qualitatives

Cette situation produisant un tableau à plus de deux lignes et plus de deux colonnes, elle conduit à des résultats ininterprétables : les pourcentages à comparer, qui sont testés par le test statistique du Chi-2, ne peuvent pas s'exprimer de façon claire et intelligible. Je ne fournirai donc aucun exemple d'une telle situation, et je vous invite plus que fortement à rendre binaire (au moins) une des deux variables lorsque vous souhaitez croiser deux variables qualitatives. Par exemple, si vous souhaitiez savoir s'il existe une association entre la race (en 4 classes, avec la variable RACE_4CL) et la cholestérolémie (en 3 classes, avec la variable CHOLES_3CL), il aurait fallu soit recoder la variable RACE_4CL en une variable

[7] https://biostatgv.sentiweb.fr/?module=tests/fisher.

binaire, soit recoder la variable CHOLES_3CL en une variable binaire (soit bien entendu recoder de façon binaire ces deux variables !).

B. Croisement d'une variable binaire ou qualitative avec une variable quantitative, et tests statistiques

1. Croisement d'une variable binaire avec une variable quantitative

Nous allons prendre pour l'exemple les variables FEMELLE et ALAT. Pour savoir si ces deux variables sont associées, on clique sur « Means » (Figure 26.a), puis on sélectionne « ALAT » pour « Means of » (Figure 26.b), « FEMELLE » pour « Cross tabulate by Value of » (Figure 26.c), puis on clique sur « Ok » (Figure 26.d).

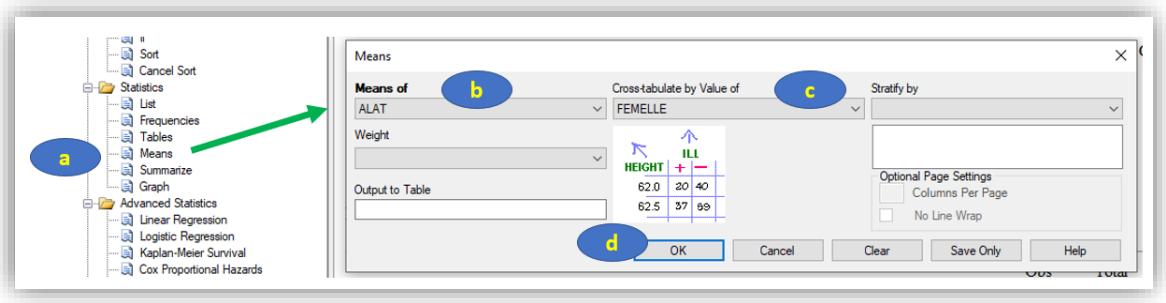


Figure 26

On obtient alors le résultat ci-dessous (cf. Figure 27). Epi Info fournit les indicateurs statistiques (moyenne, médiane, …) selon le sexe des chiens (Figure 27.a), le test de Student (Figure 27.b), le test de l'ANOVA (Figure 27.c), le test d'homogénéité des variances (Figure 27.d), et le test de Mann-Whitney / Kruskal-Wallis (Figure 27.e). Ici, l'ANOVA n'est pas pertinente puisque nous n'avons que deux moyennes ou médianes à comparer.

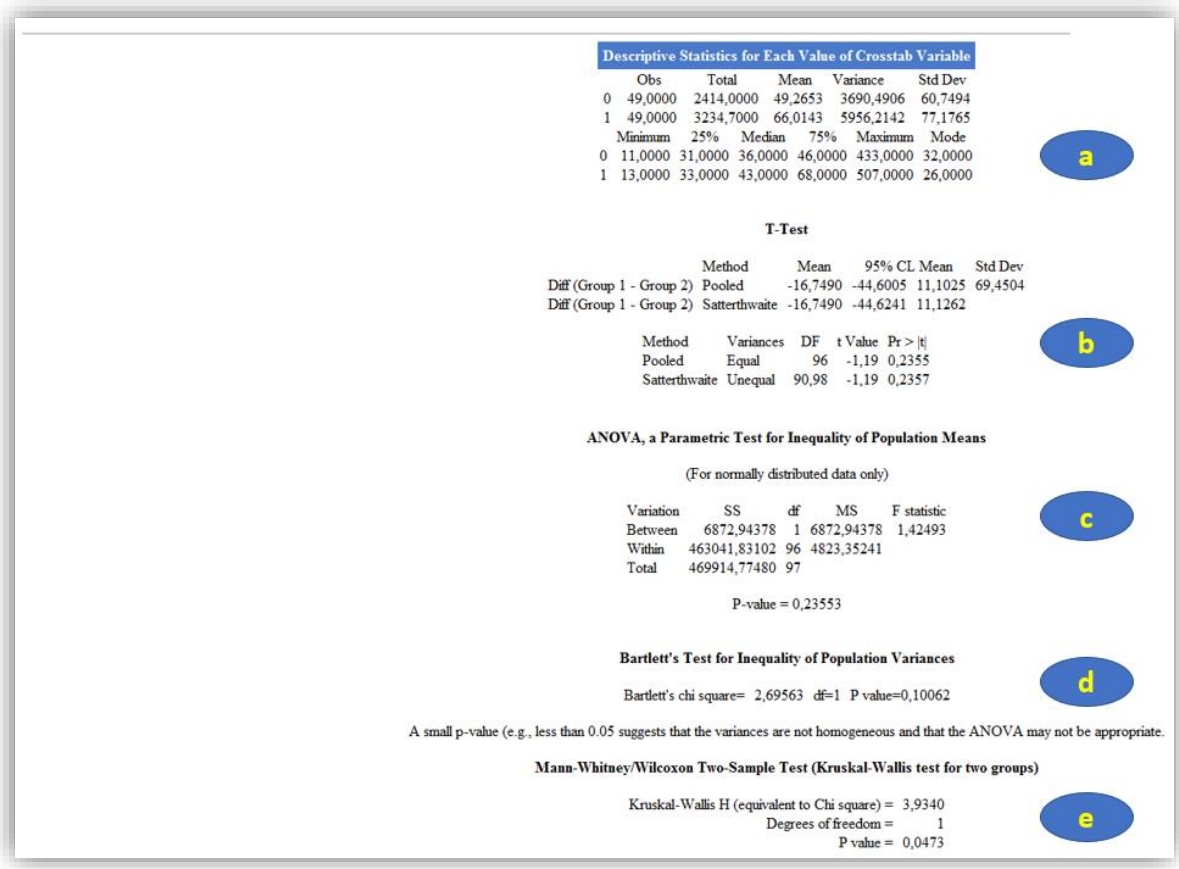


Figure 27

Je vais donc me focaliser sur les sorties (a), (b), (d), et (e) de la Figure 27 (cf. Figure 28). Les premiers résultats sont fournis sur deux lignes : pour les chiens mâles (« 0 » pour la variable FEMELLE) et pour les chiens femelles (« 1 » pour la variable FEMELLE ; Figure 28.a). Ainsi, on peut lire la moyenne de la variable ALAT chez les mâles (49,2653 UI/L) et chez les femelles (66,0143 UI/L ; Figure 28.b), ce qui conduit à une différence de moyennes de -16,7490 UI/L (Figure 28.c). (Souvenez-vous cependant qu'une moyenne d'une variable quantitative ne peut être fournie que si cette variable suit une loi normale.) Les médianes d'ALAT sont respectivement de 36,0 UI/L et 43,0 UI/L pour les mâles et les femelles (Figure 28.d). Si la distribution de la variable quantitative (ici, ALAT) suit une loi normale, alors il est possible de tester les deux moyennes avec le test de Student. Epi Info fournit le test selon que les variances peuvent être considérées comme égales ou différentes : le test de Bartlett (Figure 28.e). Notez que ce test statistique ne doit être utilisé que si la distribution de la variable quantitative suit une loi normale (si tel n'est pas le cas, comparez des médianes, et utilisez le test de Mann-Whitney !). Une règle de décision pour savoir si les variances peuvent être considérées comme égales ou différentes est de regarder le degré de signification du test de Bartlett. Ici, ce dernier vaut 0,1006. Il est supérieur à 0,05, donc on n'a pas de preuves fortes pour penser qu'elles sont réellement différentes, et l'on pourrait par conséquent les considérer comme voisines. Ainsi, dans le test de Student, on lit le degré de signification sur la ligne « Pooled » pour « Method » : 0,2355 (Figure 28.f). Si les variances n'avaient pas pu être considérées comme voisines, on aurait lu le degré de signification du test de Student sur la ligne « Satterhwaite » pour « Method » (de valeur 0,2357 ; Figure 28.f). Si pour X raison, vous préférez tester statistiquement les deux médianes d'ALAT (notamment si la variable ALAT ne suit pas une loi normale), alors vous devez utiliser le test de Mann-Whitney, dont le degré de signification se lit tout en bas de la sortie : 0,0473 (Figure 28.g). On peut remarquer que les deux moyennes ne sont pas significativement différentes tandis que les deux médianes le sont. L'une des raisons possibles de cette différence de significativité pourrait être que la variable ALAT ne suit pas

une loi normale, et l'utilisation du test de Student n'est alors pas adaptée pour savoir s'il existe une association entre les variables FEMELLE et ALAT.

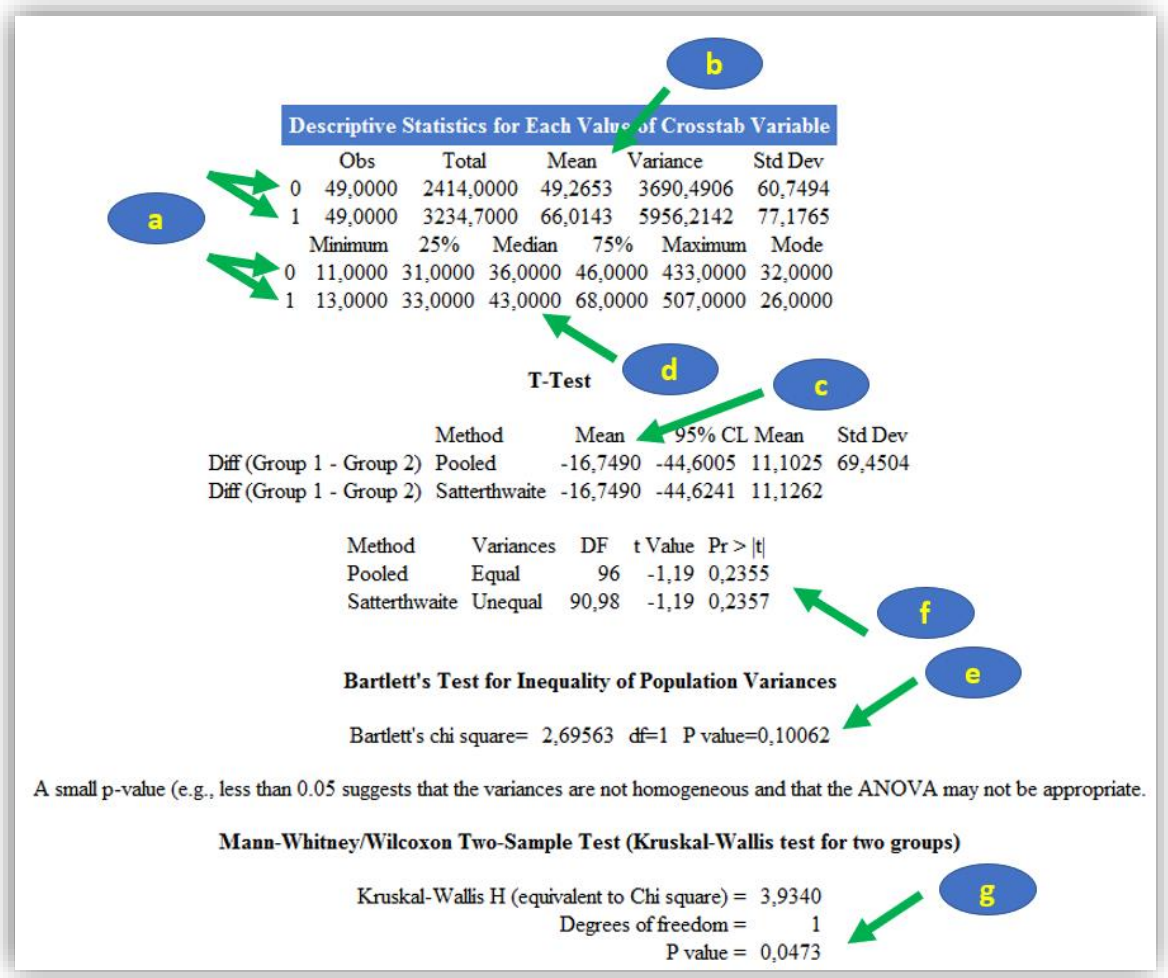


Figure 28

Notez que si nous avions utilisé la commande « Stratify by » plutôt que « Cross tabulate by value of » (Figure 29.a), nous aurions obtenu les indicateurs statistiques pour une variable quantitative classiques (moyenne, médiane, … ; Figure 29.b), bien séparés entre les femelles et les mâles, mais sans test statistique testant les indicateurs (test de Student ou de Mann-Whitney).

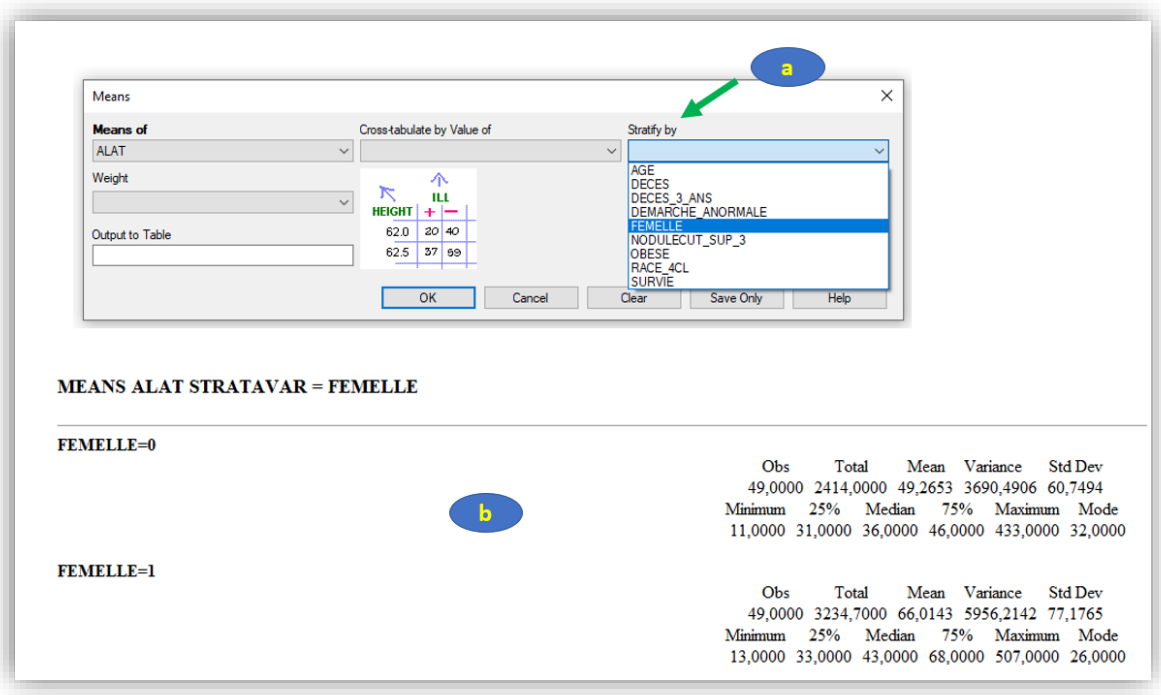


Figure 29

2. *Croisement d'une variable qualitative avec une variable quantitative*

Nous allons prendre pour l'exemple les variables RACE_4CL et ALAT. Pour savoir si ces deux variables sont associées, on sélectionne les deux variables RACE_4CL et ALAT dans la fenêtre qui s'ouvre après avoir cliqué sur « Means » (cf. démarche illustrée sur la Figure 26), on obtient les résultats présentés sur la Figure 30. Epi Info fournit les quatre moyennes (Figure 30.a) et les quatre médianes (Figure 30.b), une par modalité de la variable RACE_4CL. Epi Info indique que les variances ne peuvent pas être considérées comme égales (degré de signification du test de Bartlett inférieure à 0,0001 ; Figure 30.c), donc le degré de signification issu du test de l'ANOVA comparant les quatre moyennes, de valeur 0,0380 (Figure 30.d) ne doit pas être lu. Ce sont donc les quatre médianes qui doivent être comparées et testées, notamment pas le test de Kruskal-Wallis. Le degré de signification de ce test a pour valeur 0,0017 (Figure 30.e). Les quatre médianes sont donc significativement différentes.

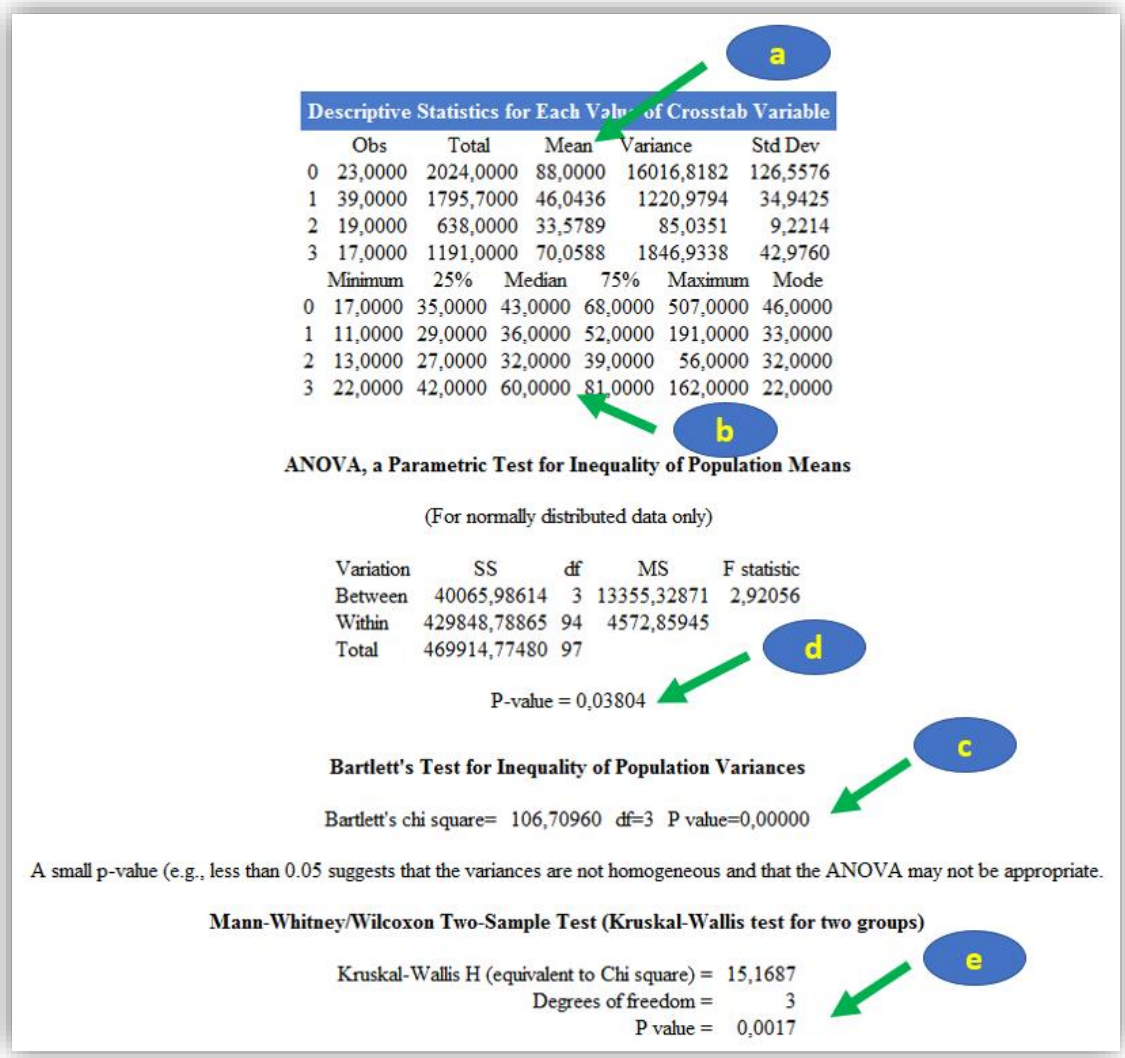


Figure 30

IV. Travailler dans un sous échantillon

Il peut être parfois intéressant de travailler sur un sous-échantillon de l'échantillon initial, en réalisant une sélection sur la valeur d'une ou plusieurs variables. Supposons que l'on veuille tout d'abord sélectionner seulement les chiens femelles de l'échantillon. Pour cela, on clique sur « Select » (cf. Figure 31.a), on sélectionne la variable sur laquelle on veut sélectionner les individus (ici, FEMELLE ; Figure 31.b), on tape la formule « = 1 » pour sélectionner les individus voulus (puisque l'on ne veut que les chiens femelles, les individus sélectionnés seront les chiens pour lesquels la variable FEMELLE vaudra « 1 » ; Figure 31.c), puis on clique sur « Ok » (Figure 31.d). Notez qu'au lieu de sélectionner la variable FEMELLE dans la liste déroulante (Figure 31.b), on aurait pu directement taper la formule « FEMELLE = 1 » dans le champ (Figure 31.c).

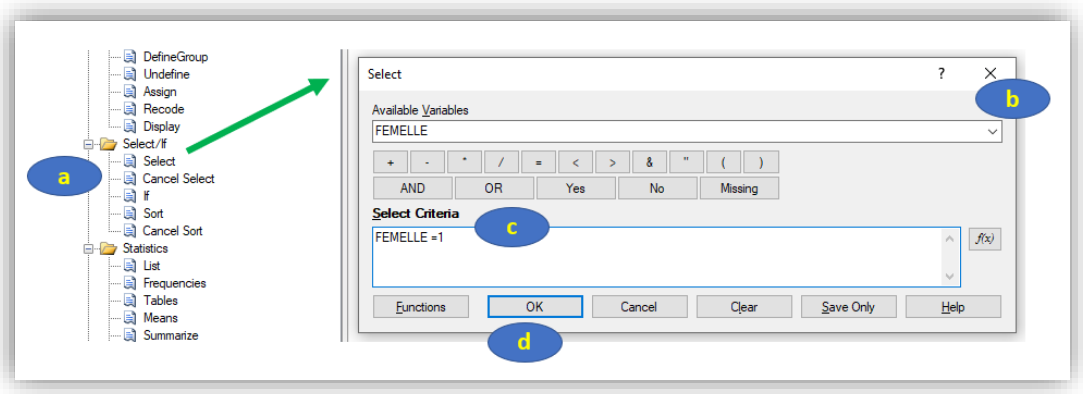


Figure 31

Epi Info nous indique le résultat de la sélection (Figure 32). Parmi les 99 chiens de l'échantillon initial, la sélection des seuls chiens femelles conduit à une taille d'échantillon de 49 chiens (femelles). Attention, toutes les statistiques que l'on réalisera à partir de là ne seront réalisées *que* sur ces 49 chiennes.

Figure 32

Pour annuler cette sélection et revenir à l'échantillon initial, il faut cliquer sur « Cancel Select » (Figure 33.a), puis sur « Ok » (Figure 33.b).

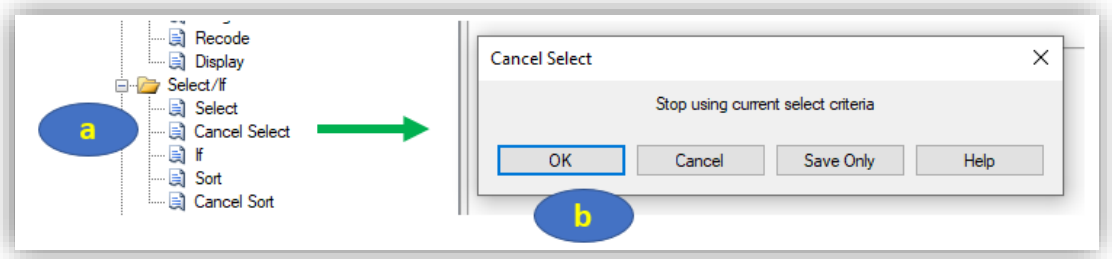


Figure 33

Supposons maintenant que l'on souhaite sélectionner, à partir de l'échantillon initial, les chiens mâles de plus de 9 ans, on tape alors la formule « FEMELLE = 0 AND AGE > 9 » dans le champ qui apparait après avoir cliqué sur « Select » (Figure 34.a), puis on clique sur « Ok » (Figure 34.b).

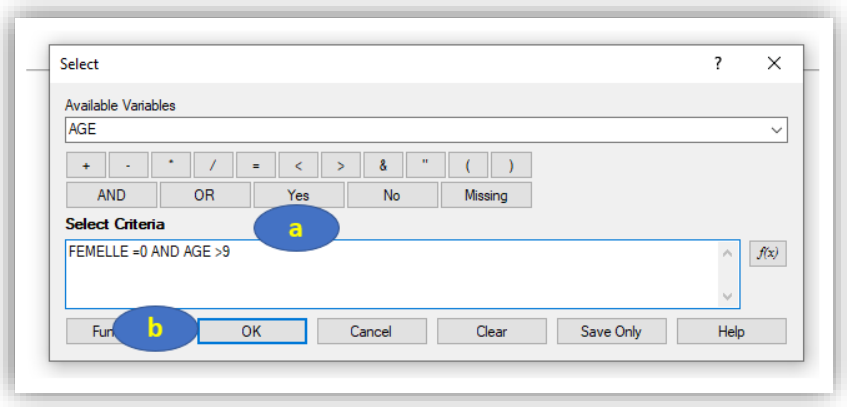


Figure 34

V. Analyse de survie à l'aide des courbes de Kaplan-Meier

A. Courbe de survie globale dans l'ensemble de l'échantillon

Avant de vous lancer dans l'analyse de survie, n'hésitez pas à (re)lire le polycopié d'analyse de survie. Pour dresser la courbe de survie globale, c'est-à-dire une seule courbe de survie pour l'ensemble de l'échantillon afin de décrire l'incidence de l'événement considéré, vous devez au préalable créer dans votre base de données une variable constante, qui prend une seule valeur pour tous les individus de l'échantillon. C'est ainsi que j'ai dû créer la variable CONSTANTE qui vaut « 1 » pour tous les chiens de l'échantillon. Dans tous les exemples qui vont suivre dans cette partie « analyse de survie de Kaplan-Meier », je vais utiliser comme événement d'intérêt la survenue d'un décès au cours du temps (variable DECES, qui vaut « 1 » pour les chiens décédés au cours de l'étude, et « 0 » pour les chiens censurés). Le temps de survie a été créé dans la base de données (variable SURVIE).

Pour réaliser une courbe de survie globale dans Epi Info, on clique sur « Kaplan-Meier Survival » (Figure 35.a), on sélectionne la variable relative à l'événement (DECES ici ; Figure 35.b), la valeur que prend cette variable pour les individus qui présentent l'événement (ici « 1 »[8] ; Figure 35.c), et la variable relative au temps de survie (ici SURVIE ; Figure 35.d). Pour faire apparaître une seule courbe de survie pour l'ensemble des individus, on doit sélectionner la variable CONSTANTE sous « Test Group Variable » (Figure 35.e). On laisse ensuite la sélection « Survival Probability » par défaut pour « Graph Type » (Figure 35.f). Cela dit, les autres propositions sous « Graph Type » donnent malheureusement exactement la même sortie que « Survival Probability » (du moins, sur mon ordinateur). Puis on clique sur « Ok » (Figure 35.g).

[8] Et je vous recommande de systématiquement coder, pour une variable relative à l'événement dans une analyse de survie, « 0 » pour les individus censurés, et « 1 » pour les individus ayant présenté l'événement.

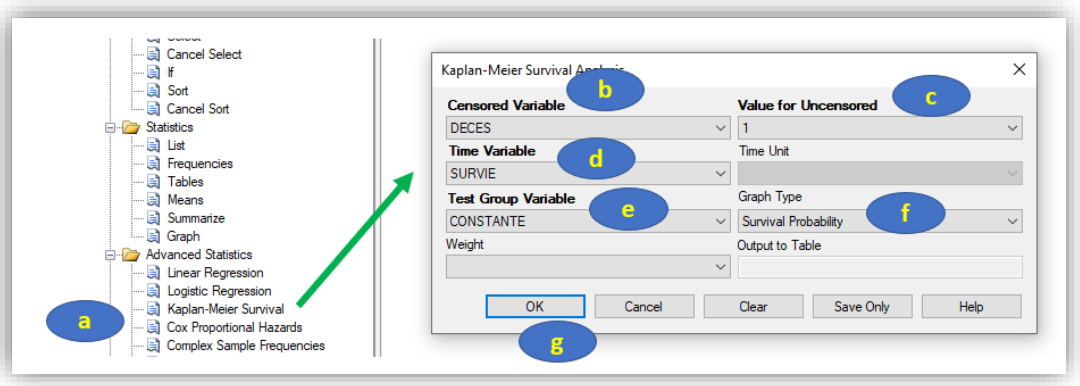


Figure 35

On obtient la courbe de survie de Kaplan-Meier ci-dessous (Figure 36).

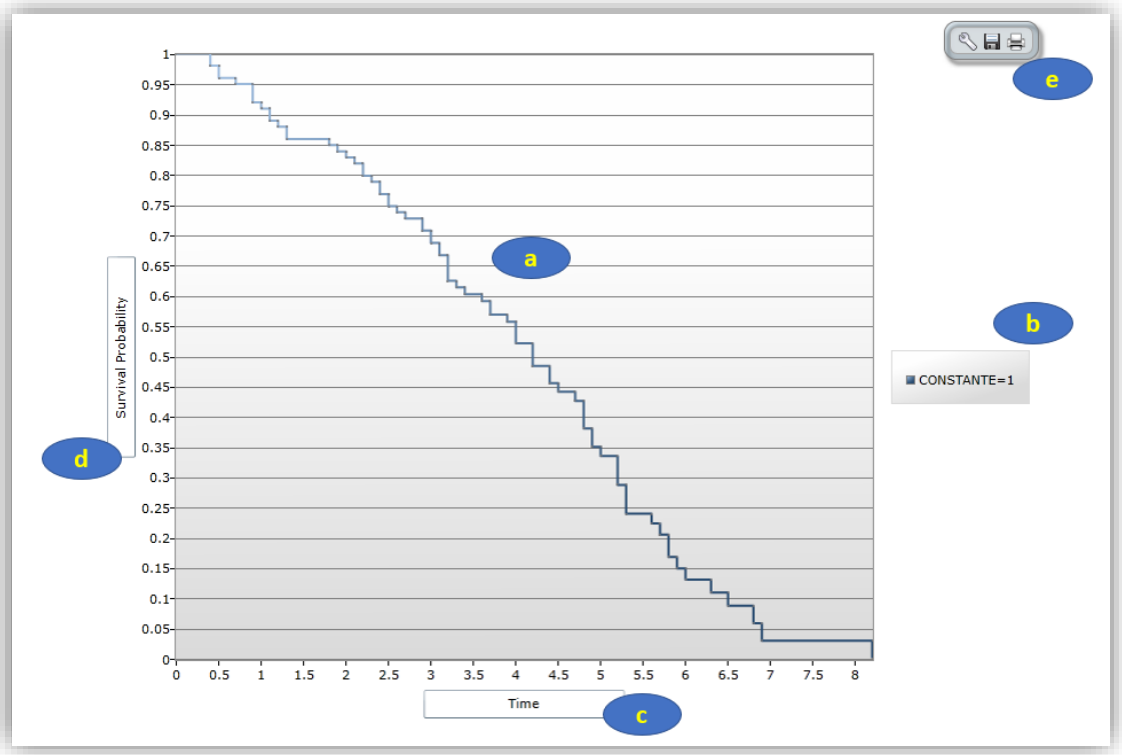


Figure 36

Puisqu'il s'agit d'une courbe de Kaplan-Meier globale pour l'ensemble de l'échantillon, il n'y a qu'une seule courbe de survie (Figure 36.a). Epi Info précise la variable utilisée ainsi que les valeurs des différentes modalités (Figure 36.b). Les titres des axes des abscisses et des ordonnées sont respectivement « Time » (Figure 36.c) et « Survival Probability » (Figure 36.d) par défaut. Ils peuvent être modifiés en cliquant sur la clé en haut à droite de la figure (Figure 36.e).

B. Courbes de survie selon les modalités d'une variable binaire

Tout d'abord, vous devez vous souvenir que l'on ne peut pas utiliser les courbes de Kaplan-Meier avec une variable quantitative. La variable doit être forcément binaire ou qualitative. Supposons que l'on veuille savoir si la présence d'une démarche anormale observée lors de la consultation chez le vétérinaire est associée au décès chez les chiens de l'étude. Pour cela, nous allons utiliser la variable

DEMARCHE_ANORMALE. On fait exactement comme précédemment (cf. Figure 35), sauf que l'on sélectionne la variable DEMARCHE_ANORMALE au lieu de la variable CONSTANTE (Figure 37.a).

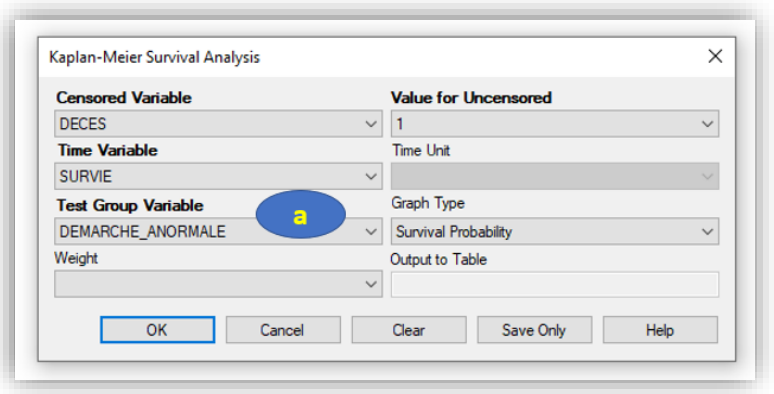


Figure 37

On obtient les deux courbes de Kaplan-Meier ci-dessous (Figure 38).

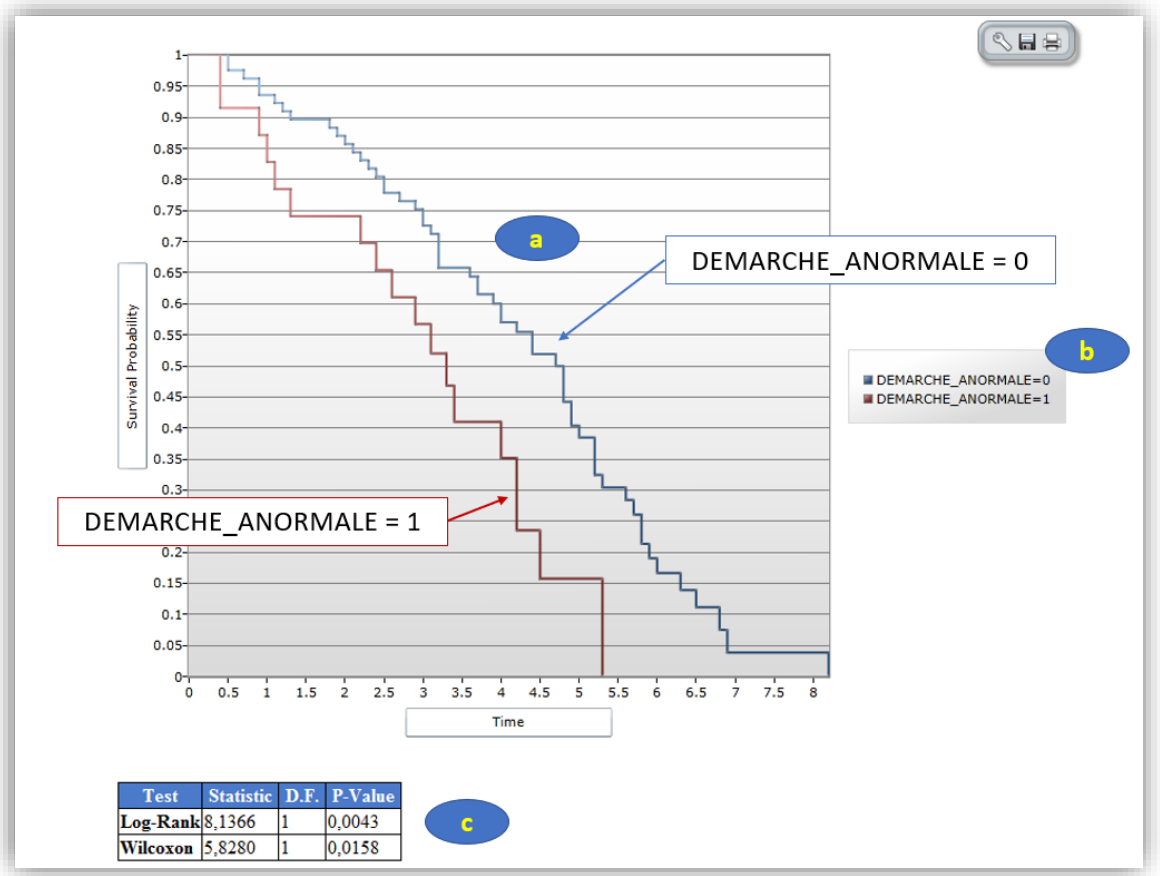


Figure 38

Les chiens représentés par la courbe rouge (courbe « DEMARCHE_ANORMALE = 1 ») décèdent plus rapidement que ce représentés par la courbe bleue (courbe « DEMARCHE_ANORMALE = 0 » ; Figure 38.a). Grâce à la légende (Figure 38.b), on voit que la courbe rouge représente les chiens dont la démarche est anormale (DEMARCHE_ANORMALE = 1) et la courbe bleue représente les chiens dont la démarche est normale (DEMARCHE_ANORMALE = 0). Cette différence de survenue d'événement est

d'ailleurs significative, car le degré de signification du test du log-rank (de valeur 0,0043 ; Figure 38.c) est inférieur à 0,05.

C. Courbes de survie selon les modalités d'une variable qualitative

Supposons que l'on veuille savoir si l'âge est associé à la survenue d'un décès chez les chiens de l'étude. L'âge étant quantitatif, il n'est pas possible d'utiliser la variable AGE. Il faut la rendre qualitative. Nous allons utiliser la variable AGE_4CL, codée en 0/1/2/3 à partir des quartiles de la variable AGE. Nous faisons exactement comme précédemment, en sélectionnant AGE_4CL (Figure 39.a).

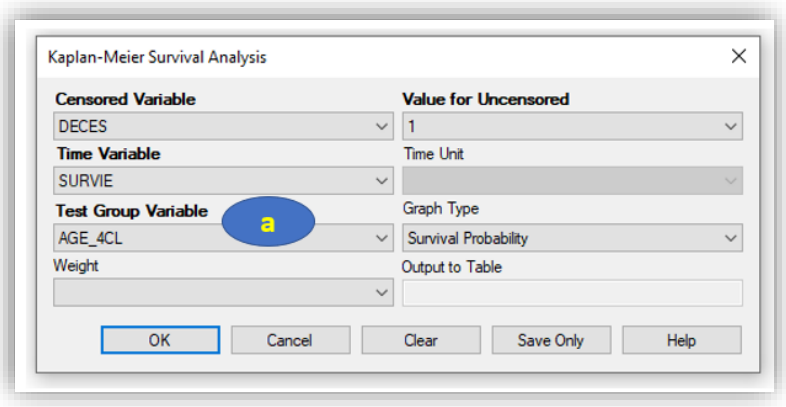


Figure 39

On obtient les quatre courbes de Kaplan-Meier ci-dessous (Figure 40).

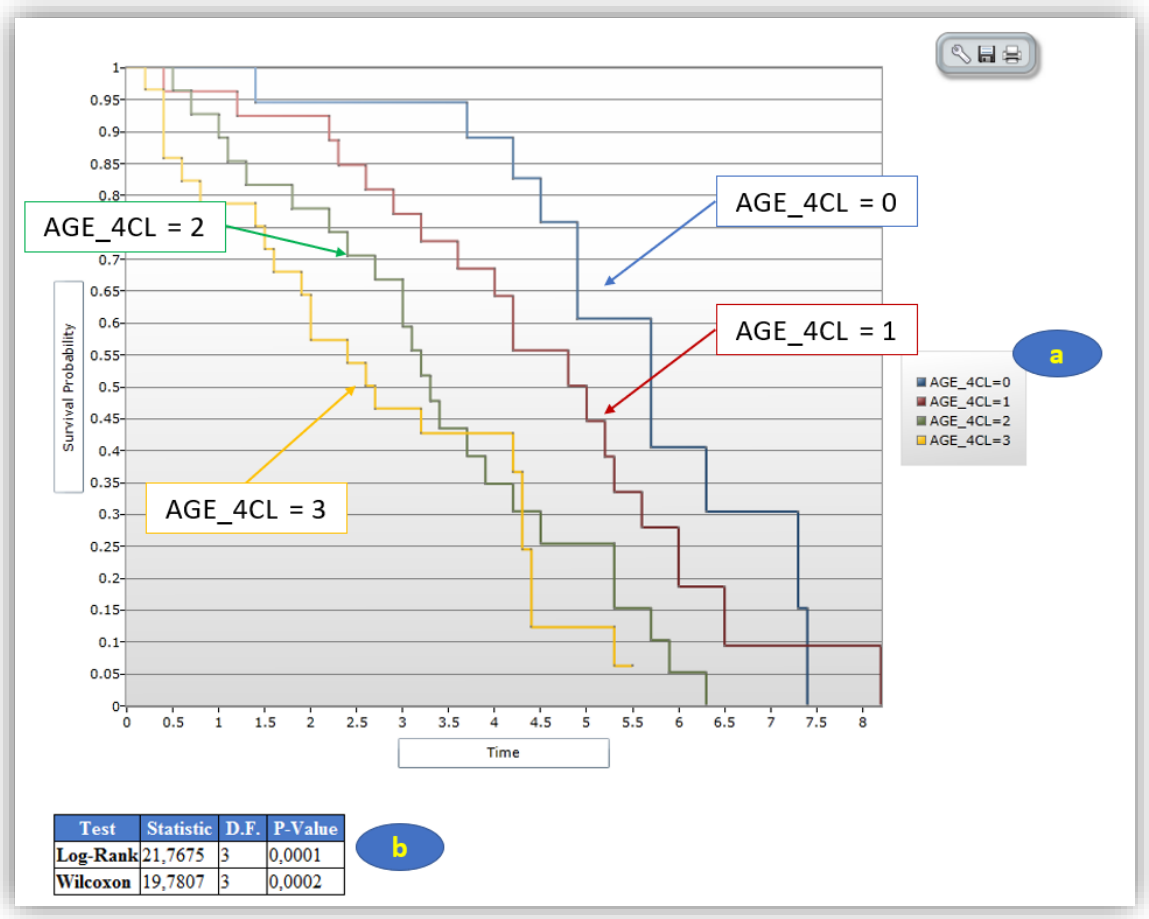


Figure 40

Grâce à la légende de la figure (Figure 40.a), on remarque que les chiens de moins de 7 ans (AGE_4CL = 0) sont moins rapidement décédés que les chiens dont l'âge est compris entre 7 et 9 ans ((AGE_4CL = 1), eux-mêmes étant moins rapidement décédés que les chiens dont l'âge est compris entre 9 et 11 ans (AGE_4CL = 2), eux-mêmes étant moins rapidement décédés que les chiens d'âge supérieur ou égal à 11 ans (AGE_4CL = 3). En effet, la courbe de survie bleue est au-dessus de la courbe de survie rouge, elle-même étant au-dessus de la courbe de survie verte, elle-même étant au-dessus de la courbe de survie jaune. Cette différence de survenue d'événement est d'ailleurs significative, car le degré de signification du test du log-rank (de valeur 0,0001 ; Figure 40.b) est inférieur à 0,05. Attention à ne pas sur-interpréter cette significativité. Bien que l'on *observe* une relation « dose-effet » avec l'âge (plus la modalité de l'âge augmente, et plus la survenue de décès était rapide), le test du log-rank ne teste pas de relation dose-effet. Ainsi, il n'est pas question de dire que la survenue de décès était *significativement* plus rapide lorsque l'âge à J0 augmentait (cf. commentaire identique sur la significativité d'un test ANOVA dans le polycopié de bases en biostatistique).

Chapitre 3 – Modèles de régression

I. Introduction

A. Gestion du symbole de la décimale des variables quantitatives

Si vous souhaitez inclure dans un modèle de régression une variable qui comporte un chiffre après une virgule pour certains individus, alors il semble malheureusement nécessaire (en tout cas, je n'ai pas trouvé d'autre solution) de paramétrer votre ordinateur de telle façon que le séparateur de décimal soit le point (« . ») et non la virgule (« , »). Pour cela, on clique sur les paramètres Windows, on cherche le panneau de configuration (Figure 41.a), on clique sur « Modifier les formats de date » (Figure 41.b), puis sur « Paramètres supplémentaires… » (Figure 41.c), on tape le « . » comme symbole décimal (Figure 41.d), puis on clique sur « Ok » (Figure 41.e).

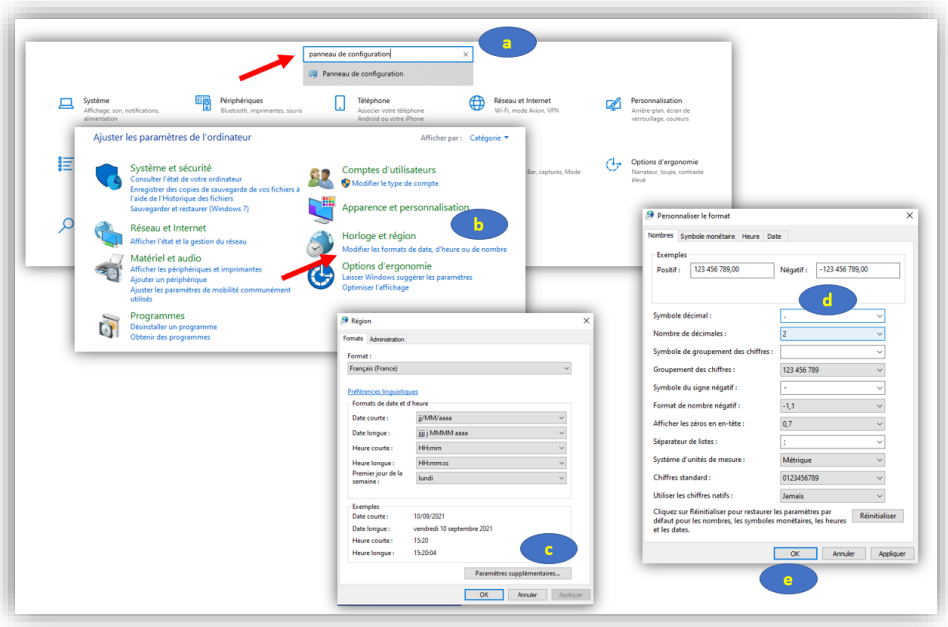


Figure 41

II. Théorie des modèles de régression

A. Ecriture d'un modèle de régression

Un modèle de régression met en relation le CdJ, quantifié par Y, et une ou plusieurs expositions (⇔ variables) E. Tout d'abord, il est indispensable que chacune des variables incluses dans un modèle soit une variable numérique (cf. Chapitre 2, Partie I.A, page 15). Ainsi, un modèle de régression comprenant N variables E_i ($i \in \{1, \ldots, n\}$) s'écrit de façon générale :

$$\bar{Y}_{/E_1,E_2,\ldots,E_N} = \alpha + \sum_{i=1}^{N} \beta_i . E_i$$

En français, ce « $\bar{Y}_{/E_1,E_2,\ldots,E_N}$ » se lit « la valeur de l'espérance de Y sachant les valeurs des variables E_1, E_2, …, E_N ». Le mot « espérance », qui est un terme mathématique[9], peut être compris comme « la valeur attendue de Y à partir des valeurs des variables incluses dans le modèle ».

Si N = 1, on dira que le modèle de régression est « univarié » (il ne contient qu'une seule variable), et si N ≥ 2, alors on dira que le modèle de régression est multivarié. Dans le modèle, α et les β_i sont les coefficients du modèle, et E_i ($i \in \{1, \ldots, N\}$) sont les N variables que l'on souhaite inclure dans le modèle.

B. Choix d'un modèle de régression et écriture mathématique du modèle

Ce qui guide le choix d'un modèle de régression est le type de la variable relative au CdJ.

Si le CdJ est quantitatif (par exemple, la concentration en ALAT), le modèle de régression est la régression linéaire et Y est directement la variable relative au CdJ. Notez que pour utiliser un modèle de régression linéaire, le CdJ quantitatif Y doit suivre à peu près une loi normale. Si tel n'est pas le cas, je vous recommande alors de transformer cette variable quantitative en une variable binaire, en utilisant un seuil qui a un sens clinique (et vous utiliserez alors un modèle de régression logistique – cf. ci-dessous). Le modèle de régression linéaire s'écrit :

$\overline{CdJ}_{/E_1,E_2,\ldots,E_N} = \alpha + \sum_{i=1}^{N} \beta_i . E_i$, où « $\overline{CdJ}_{/E_1,E_2,\ldots,E_N}$ » est l'espérance de la valeur du CdJ quantitatif en fonction des valeurs des variables E_i incluses dans le modèle. De façon générale, β_i quantifie l'association entre le CdJ (quantitatif) et E_i en tant que valeur de la différence de moyennes du CdJ quantitatif.

Si le CdJ est binaire, et non assorti d'un temps de survenue (par exemple, dans une étude cas-témoins ou transversale), alors le modèle de régression est la régression logistique. Le modèle de régression logistique s'écrit :

$Logit\left(\bar{P}_{/E_1,E_2,\ldots,E_N}\right) = Ln\left(\frac{\bar{P}_{/E_1,E_2,\ldots,E_N}}{1-\bar{P}_{/E_1,E_2,\ldots,E_N}}\right) = \alpha + \sum_{i=1}^{N} \beta_i . E_i$, où « $\bar{P}_{/E_1,E_2,\ldots,E_N}$ » est l'espérance de la probabilité de présenter le CdJ en fonction des valeurs des variables E_i incluses dans le modèle. De façon générale, β_i quantifie l'association entre la *présence* du CdJ (binaire) et E_i en tant que valeur du $Ln\left(OR_{E_i}\right)$, où OR_{E_i} est l'Odds Ratio quantifiant l'association entre la présence du CdJ et la variable E_i.

Si le CdJ est binaire et assorti d'un temps de survenue (par exemple, dans une étude de cohorte), alors le modèle de régression est le modèle de Cox. Le modèle de Cox s'écrit :

$Ln\left(\overline{\lambda(t)}_{/E_1,E_2,\ldots,E_N}\right) = Ln\left(\lambda_0(t)\right) + \sum_{i=1}^{N} \beta_i . E_i$, où « $\overline{\lambda(t)}_{/E_1,E_2,\ldots,E_N}$ » est l'espérance de l'incidence instantanée du CdJ en fonction de la valeur des variables E_i incluses dans le modèle. De façon générale, β_i quantifie l'association entre la *survenue* du CdJ (binaire) et E_i en tant que valeur du $Ln\left(HR_{E_i}\right)$, où HR_{E_i} est le Risque Relatif (« Hazard Ratio » pour un modèle de Cox, qui est un « rapport des incidences instantanées ») quantifiant l'association entre la survenue du CdJ et la variable E_i.

C. Problématique des données manquantes

Ce point est très important et il est souvent omis par les utilisateurs de modèles de régression. Les individus de l'échantillon qui sont utilisés pour estimer les coefficients du modèle $\bar{Y}_{/E_1,E_2,\ldots,E_N} = \alpha + \sum_{i=1}^{N} \beta_i . E_i$ sont les individus de l'échantillon tels qu'aucune de leurs *N* variables E_i n'a de donnée manquante (et bien entendu, ces individus ne doivent pas non plus avoir de donnée manquante sur Y).

[9] https://fr.wikipedia.org/wiki/Esp%C3%A9rance_math%C3%A9matique

La Figure 42 ci-dessous présente un exemple fictif d'un fichier de données de 6 individus, pour 4 variables : Y, E_1, E_2, et E_3. Dans ce fichier de données, l'individu #1 a une donnée manquante pour la variable E_2, l'individu #2 a une donnée manquante pour la variable E_3, l'individu #3 a une donnée manquante pour la variable Y, l'individu #4 a une donnée manquante pour la variable E_1, et les individus #5 et #6 n'ont aucune donnée manquante pour les 4 variables.

ID	Y	E1	E2	E3
1	5	56		55
2	6	34	11	
3		89	24	87
4	2		21	90
5	4	77	9	65
6	3	54	27	50

Figure 42

Le modèle de régression (A) $\bar{Y}_{/E_1,E_2,E_3} = \alpha + \beta_1.E_1 + \beta_2.E_2 + \beta_3.E_3$ ne tournera que sur les individus #5 et #6, car ce sont uniquement ces deux individus pour lesquels aucune donnée ne manque sur les variables E_1, E_2, E_3, et Y. Le modèle de régression (B) $\bar{Y}_{/E_1,E_2} = \alpha + \gamma_1.E_1 + \gamma_2.E_2$ tournera quant à lui sur les individus #2, #5, et #6.

Il y a deux conséquences de cela très importantes. La première, c'est qu'un modèle tourne parfois sur beaucoup moins d'individus qu'attendus, même si, individuellement, chaque individu a très peu de données manquantes. Parfois, il faudra admettre de ne pas inclure une variable dans un modèle si elle est manquante pour beaucoup d'individus (ce qui est très embêtant si cette variable est un facteur de confusion). La seconde est la suivante. Si vous souhaitez comparer les valeurs de β_2 et γ_2 (notamment pour savoir si la variable E_3 a joué un rôle de confusion dans l'étude de l'association entre le CdJ et E_2), ces valeurs de β_2 et γ_2 ne pourront être comparées que si les deux modèles (A) et (B) tournent sur les *mêmes* individus – ce qui n'est pas le cas sans intervention de votre part (puisque l'individu #2 a été utilisé pour estimer les valeurs de γ_1 et γ_2 mais cet individu #2 n'a pas été utilisé pour estimer les valeurs de β_1, de β_2, et de β_3). Pour imposer le fait que ces deux modèles (A) et (B) tournent sur les mêmes individus (les individus #5 et #6), il faut soit avoir créé au préalable un fichier de données qui ne comprend que les individus sur lesquels on veut faire tourner les différents modèles (sans aucune donnée manquante sur les variables incluses dans les modèles), soit contraindre le logiciel de faire tourner le modèle (B) seulement sur les individus #5 et #6 (cf. Chapitre 2, Partie IV, page 28).

III. La régression linéaire

A. Introduction

Même si vous ne prévoyez de n'utiliser que la régression logistique ou que le modèle de Cox, il est néanmoins indispensable de lire toute cette partie sur la régression linéaire. En effet, bien que l'interprétation des résultats d'une régression logistique ou d'un modèle de Cox ne soit pas la même que celle d'une régression linéaire, la démarche d'interprétation est quant à elle identique aux trois modèles.

Pour illustrer l'interprétation des résultats d'une régression linéaire, je vais prendre comme variable relative au CdJ la concentration en ALAT (dont on suppose qu'elle suit une loi normale).

Comme nous l'avons vu dans la partie II.B de ce Chapitre (page 36), un modèle de régression linéaire permet d'estimer la valeur de l'espérance du CdJ quantitatif Y d'un individu sachant les valeurs de ses

variables E_i incluses dans le modèle, grâce aux estimations de α et des β_i. Par exemple, faisons tourner le modèle suivant à partir des données de notre échantillon :

$$\overline{ALAT}_{/FEMELLE,AGE} = \alpha + \beta_1.FEMELLE + \beta_2.AGE$$

Pour cela avec Epi Info, on clique sur « Linear Regression » (Figure 43.a), on sélectionne la variable Y du modèle, ici « ALAT » (Figure 43.b), on sélectionne les variables que l'on veut inclure dans le modèle dans la liste déroulante sous « Other Variables » (Figure 43.c), ici les variables FEMELLE et AGE, puis on clique sur « Ok » (Figure 43.d).

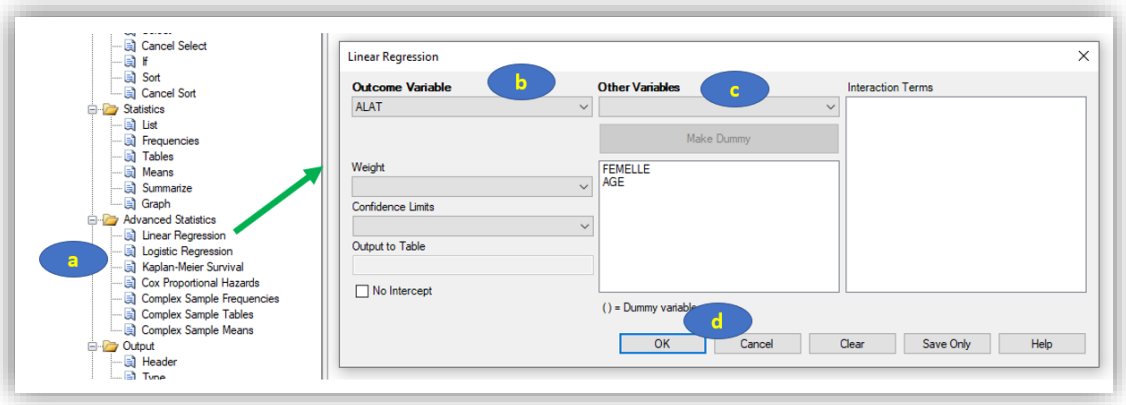


Figure 43

On obtient les résultats présentés sur la Figure 44. La colonne « Coefficient » (a) est celle dans laquelle figurent les coefficients du modèle.

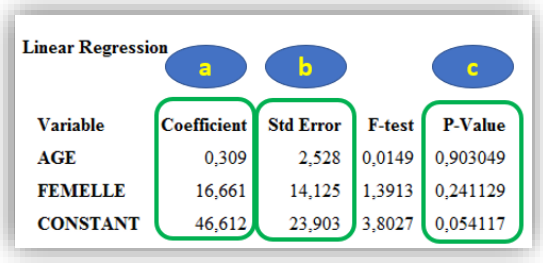


Figure 44

Ainsi, on peut voir sur la Figure 44 que le logiciel estime les valeurs de 46,612, 16,661 et 0,309 respectivement pour α, β_1, et β_2. Le modèle, estimé à partir des données de l'échantillon, s'écrit donc :

$$\overline{ALAT}_{/FEMELLE,AGE} = 46{,}612 + 16{,}661.FEMELLE + 0{,}309.AGE$$

Cela signifie qu'à partir des données de l'échantillon, le modèle estime qu'*un* chien mâle (FEMELLE = 0) de 12 ans (AGE = 12) avait, en espérance, une concentration en ALAT de : 46,612 + 16,661x0 + 0,309x12 = 50,320 UI/L. Ce qui équivalent à dire[10] que dans l'échantillon, *les* chiens mâles de 12 ans avaient une *moyenne* de concentration en ALAT de 50,320 UI/L. Le modèle estime aussi que la moyenne de la concentration en ALAT des chiens femelles de l'échantillon de 7 ans est de : 46,612 + 16,661x1 + 0,309x7 = 65,436 UI/L.

A partir des exemples ci-dessus, on peut déjà interpréter la valeur de α : c'est la valeur de $\bar{Y}_{/E_1=0,E_2=0,\ldots,E_N=0}$. Autrement dit, c'est l'espérance de Y lorsque les valeurs des variables E_i incluses

[10] Et je vous passe la démonstration 😊

dans le modèle sont toutes égales à 0. Dans l'exemple ci-dessus, voici donc l'interprétation qu'*aurait* α (de valeur 46,612) : ce serait la moyenne de la concentration en ALAT pour les chiens pour lesquels les variables FEMELLE et AGE sont toutes deux égales à 0. Or, dans la mesure où l'échantillon ne comprend certainement pas des chiots qui viennent de naitre (un chien qui vient de naître a effectivement un âge de 0,0 an), la valeur de 46,612 estimée par le modèle n'est pas interprétable. De façon plus générale, si une variable incluse dans le modèle ne prend jamais la valeur 0 parmi les individus de l'échantillon, alors la valeur de α n'est pas interprétable.

Notez que l'on n'utilise que rarement les valeurs des coefficients dans l'objectif d'estimer la valeur moyenne de Y pour des caractéristiques fixées des animaux. On utilise bien davantage ces coefficients pour quantifier l'association entre le CdJ et la variable correspondante au coefficient. Et c'est ce que nous allons voir ci-dessous, avec l'interprétation des coefficients β (je n'interpréterai plus, dans la suite de ce guide, la valeur de α).

La colonne « Std Error » (Figure 44.b) correspond, comme son nom l'indique, à la Standard Error (SE) des coefficients α (SE_α), β_1 ($SE_{\beta 1}$), et β_2 ($SE_{\beta 2}$) du modèle. En pratique, elle permet de calculer l'$IC_{95\%}$ d'un coefficient β (ainsi que du coefficient α, mais comme écrit ci-dessus, on n'accorde que très peu, voire pas, d'importance à α) : $\beta \pm 1{,}96.SE_\beta$. Ainsi, l'$IC_{95\%}$ de β_1 est le suivant : 16,661 ± 1,96x14,125, soit [-11,024 ; 44,346].

La colonne « P-Value » (Figure 44.c) indique le degré de signification testant la significativité des coefficients. Le test statistique est le test de Wald[11], il teste si le coefficient est significativement différent de 0. Sur la Figure 44, on peut donc voir que le coefficient β_1 de valeur 16,661 n'est pas significativement différent de 0 (p = 0,24).

B. Interprétation des résultats d'une régression linéaire univariée

1. *Cas général*

Le modèle de régression linéaire univarié s'écrit : $\bar{Y}_{/E} = \alpha + \beta.E$, avec E une variable quelconque (binaire, qualitative, ou quantitative). Pour interpréter la valeur de β, je vais écrire ce modèle pour deux groupes d'animaux : un groupe au sein duquel les animaux ont une valeur égale à e_1 pour E, et un second groupe au sein duquel les animaux ont une valeur égale à e_2 pour E.

$$\bar{Y}_{/E=e_1} = \alpha + \beta.e_1$$

$$\bar{Y}_{/E=e_2} = \alpha + \beta.e_2$$

Ainsi, et comme je l'ai déjà écrit ci-dessus, le modèle estime que la moyenne de Y chez les animaux dont E vaut e_1 est $\alpha + \beta.e_1$, et que la moyenne de Y chez les animaux dont E vaut e_2 est $\alpha + \beta.e_2$. Maintenant, je fais la soustraction entre les deux estimations :

$$\bar{Y}_{/E=e_2} - \bar{Y}_{/E=e_1} = (\alpha + \beta.e_2) - (\alpha + \beta.e_1) = \beta.(e_2 - e_1)$$

Ainsi, lorsque l'écart sur la valeur de la variable E entre deux groupes d'animaux vaut +1 ($\Leftrightarrow e_2 - e_1 = +1$), alors $\bar{Y}_{/E=e_2} - \bar{Y}_{/E=e_1} = \beta$.

Par conséquent, β s'interprète de la façon suivante dans un modèle de régression linéaire univarié : β est la valeur, estimée à partir des données de l'échantillon, de la différence de moyennes de Y entre deux groupes d'animaux différant de +1 unité pour leur variable E, quelles que soient les valeurs de leur variable E.

Cette interprétation ci-dessus est fondamentale. Nous allons voir les conséquences d'une telle interprétation en fonction des différents types de variable (binaire, qualitatif, et quantitatif).

[11] https://en.wikipedia.org/wiki/Wald_test

Par ailleurs, si β = 0, cela signifie qu'il n'existe aucune différence de moyennes de Y entre deux groupes d'animaux différant de +1 unité pour leur variable E. Ainsi, β = 0 traduit une absence totale d'association entre le CdJ (quantifié par Y) et E. C'est la raison pour laquelle le test statistique de Wald testant la significativité de β teste si β est significativement différent de « 0 ». Si tel est le cas, alors il existe une association significative entre le CdJ (quantifié par Y) et E dans l'échantillon.

2. *Modèle de régression linéaire univarié avec variable binaire*

Supposons le modèle de régression linéaire suivant, incluant une seule variable binaire E :

$$\bar{Y}_{/E} = \alpha + \beta.E$$

Je recommande fortement le codage d'une variable binaire dans le fichier de données de telle façon à ce que les animaux exposés à E aient une valeur pour E égale à 1, et à ce que les animaux non exposés à E aient une valeur pour E égale à 0. Ainsi, β est la valeur, estimée à partir des données de l'échantillon, de la différence de moyennes de Y entre les animaux exposés et les animaux non exposés (car avec le codage fortement recommandé, $e_2 - e_1 = +1$).

Par exemple, faisons tourner le modèle suivant :

$$\overline{ALAT}_{/FEMELLE} = \alpha + \beta.FEMELLE$$

Pour faire tourner ce modèle dans Epi Info, après avoir cliqué sur « Linear Regression » (Figure 43.a), on sélectionne la variable Y du modèle, ici ALAT (Figure 45.a), on sélectionne la variable FEMELLE (Figure 45.b), puis on clique sur « Ok » (Figure 45.c).

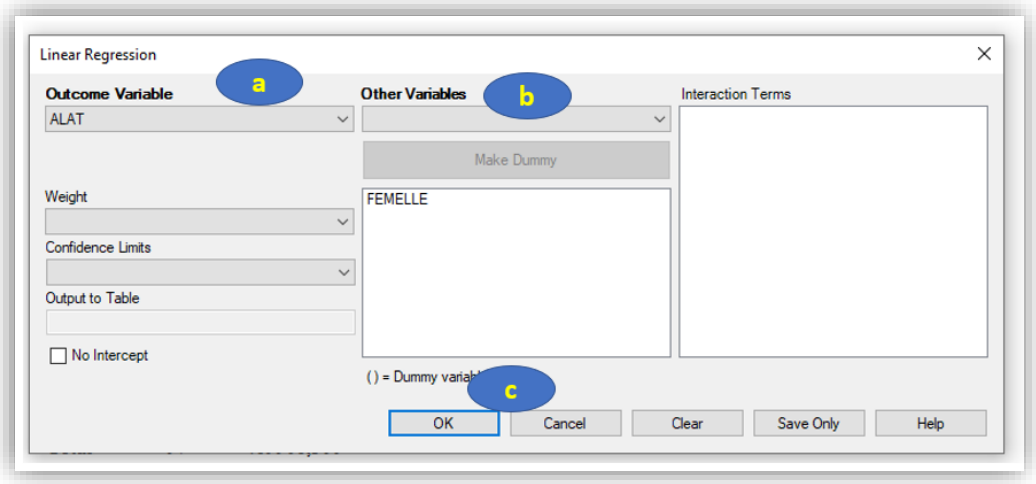


Figure 45

On obtient les résultats présentés sur la Figure 46 ci-dessous.

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
FEMELLE	16,755	14,031	1,4260	0,235358
CONSTANT	49,265	9,921	24,6571	0,000003

Figure 46

A partir des résultats de la régression linéaire présentés sur la Figure 46, le modèle de régression linéaire estimé par Epi Info reliant la concentration en ALAT au sexe des chiens s'écrit de la façon suivante :

$$\overline{ALAT}_{/FEMELLE} = 49{,}265 + 16{,}755.FEMELLE$$

La valeur de β de « 16,755 » s'interprète de la façon suivante : la différence entre la moyenne de la concentration en ALAT des chiens pour lesquelles la variable FEMELLE vaut « 1 » (⇔ les chiens femelles) et la moyenne de la concentration en ALAT des chiens pour lesquelles la variable FEMELLE vaut « 0 » (⇔ les chiens mâles) vaut 16,755 UI/L. L'$IC_{95\%}$ de la différence de moyennes de 16,755 UI/L entre les chiens femelles et les chiens mâles est : 16,755 ± 1,96x14,031, soit $[-10{,}746 ; 44{,}256]_{95\%}$. Cet $IC_{95\%}$ de la différence de moyennes ne comprenant pas « 0 », les deux moyennes ne sont pas significativement différentes. On retrouve d'ailleurs le degré de signification du test statistique testant si cette valeur de 16,755 est significativement différente de « 0 », de valeur 0,24, donc supérieure à 0,05 (cf. colonne « P-value » sur la Figure 46). Ainsi, il n'existait pas d'association significative dans l'échantillon entre la concentration en ALAT et le sexe des chiens. N'oubliez pas que les résultats de cette régression linéaire ne sont valides que si la distribution de Y (ici, la concentration en ALAT) peut être considérée comme normale.

3. *Modèle de régression linéaire univarié avec variable quantitative*

Supposons le modèle de régression linéaire suivant, incluant une seule variable quantitative E :

$$\bar{Y}_{/E} = \alpha + \beta.E$$

Nous allons voir ci-dessous que ce modèle repose sur une hypothèse que l'on appelle « l'hypothèse de la linéarité de l'association entre le CdJ (quantifié par Y) et la variable E ». Si cette hypothèse est vérifiée, le modèle est valide, et l'interprétation du coefficient β est possible. Si cette hypothèse n'est pas vérifiée, alors ce modèle ne doit pas être utilisé, car l'estimation de β ne sera pas interprétable. Pour illustrer cette hypothèse de la linéarité de l'association entre Y et E, faisons tourner le modèle suivant à partir des données de l'échantillon :

$$\overline{ALAT}_{/AGE} = \alpha + \beta.AGE$$

Pour faire tourner ce modèle dans Epi Info, après avoir cliqué sur « Linear Regression », on sélectionne la variable Y du modèle, ici la variable ALAT (a), on sélectionne la variable AGE (b), puis on clique sur « Ok » (c).

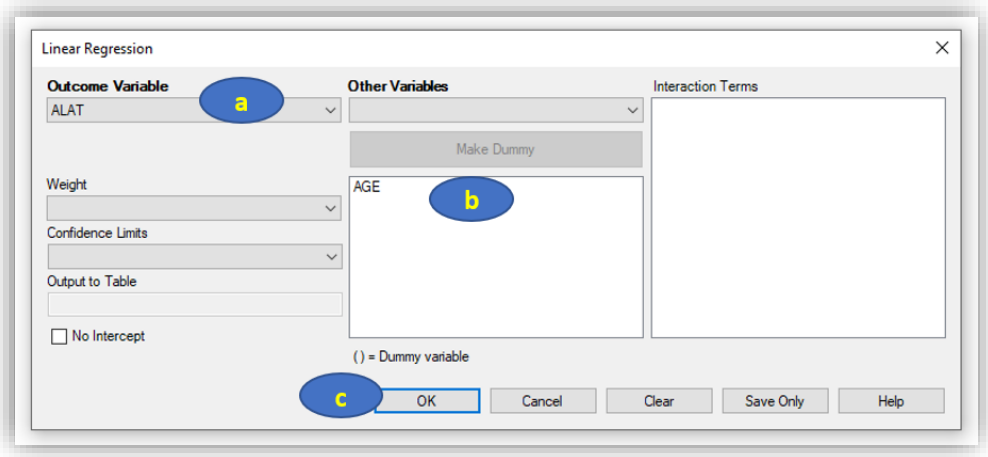
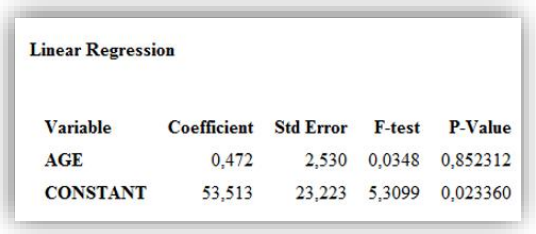


Figure 47

On obtient les résultats présentés sur la Figure 48.



Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
AGE	0,472	2,530	0,0348	0,852312
CONSTANT	53,513	23,223	5,3099	0,023360

Figure 48

A partir des résultats de la régression linéaire présentés sur la Figure 48, le modèle de régression linéaire estimé par Epi Info reliant la concentration en ALAT à l'âge des chiens s'écrit de la façon suivante :

$$\overline{ALAT}_{/AGE} = 53{,}513 + 0{,}472.AGE$$

Cela signifie qu'à partir des données de l'échantillon, le modèle estime que la différence de moyennes de la concentration en ALAT entre deux groupes d'animaux différant de +1 année d'âge est de 0,472 UI/L. Ainsi, le modèle estime qu'en moyenne, et dans l'échantillon, la différence de concentration en ALAT entre des chiens de (par exemple) 5 ans et des chiens de 4 ans est de 0,472 UI/L. Le modèle estime aussi que la différence de concentration en ALAT entre des chiens de (par exemple) 14 ans et des chiens de 13 ans est *aussi* de 0,472 UI/L. Ce modèle fait donc l'hypothèse qu'une augmentation de +1 unité sur l'âge (⇔ 1 année), quelle que soit la valeur de l'âge, se traduit par la *même* différence sur la moyenne en ALAT : c'est l'hypothèse de la linéarité de l'association entre Y et E (ici entre la concentration en ALAT et l'âge). Si en vrai, la différence de moyenne de concentration en ALAT n'est pas la même, pour une même augmentation +1 année d'âge, entre des valeurs faibles de l'âge et des valeurs de l'âge plus élevées, alors l'hypothèse de la linéarité de l'association n'est pas vérifiée, et le modèle fournira une estimation de β ininterprétable.

(Si ce que j'ai écrit ci-dessus est du charabia, relisez une seconde, troisième, … fois jusqu'à ce que ce que je viens d'écrire ne le soit plus. Si ça le reste, alors je vous recommande fortement de n'utiliser *que* des variables binaires dans vos analyses, les interprétations sont en effet beaucoup plus faciles.)

Cette hypothèse de la linéarité de l'association entre Y et E doit être systématiquement vérifiée *avant* d'inclure une variable quantitative E dans un modèle de régression (linéaire). La vérification de cette hypothèse fait l'objet de la Partie IV de ce Chapitre.

Attention, si le modèle estime une valeur de coefficient β qui est ininterprétable parce que le modèle est incorrect, ce n'est pas à cause du logiciel ou de la statistique, c'est à cause de celle ou celui qui a choisi de faire tourner un tel modèle ! C'est pour cela que conduire des analyses statistiques nécessite de savoir toutes les conditions ou hypothèses sur lesquelles reposent ces analyses. N'attendez pas qu'un logiciel vous dise « attention, vous ne devriez pas faire tourner ce modèle, car il repose sur une hypothèse qui n'a pas l'air de tenir la route, physiopathologiquement parlant » 😉.

4. Modèle de régression linéaire univarié avec variable qualitative ordinale

Supposons le modèle de régression linéaire suivant, incluant une seule variable qualitative ordinale E :

$$\bar{Y}_{/E} = \alpha + \beta.E$$

Nous allons voir ci-dessous que ce modèle, lui aussi, repose sur l'hypothèse de la linéarité de l'association entre le CdJ (quantifié par Y) et E. Là encore bien entendu, si cette hypothèse n'est pas vérifiée, alors ce modèle ne devra pas être utilisé, car l'estimation de β ne sera pas interprétable. Nous allons illustrer cela à partir du modèle suivant que nous allons faire tourner à partir des données de l'échantillon :

$$\overline{ALAT}_{/CHOLES\_3CL} = \alpha + \beta.CHOLES\_3CL$$

Après avoir sélectionné la variable Y du modèle, ici la variable ALAT, et la variable CHOLES_3CL (comme nous l'avons fait précédemment pour les variables FEMELLE et AGE), nous obtenons les résultats présentés sur la Figure 49.

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
CHOLES_3CL	8,198	9,969	0,6762	0,412944
CONSTANT	49,696	11,958	17,2720	0,000070

Figure 49

A partir des résultats de la régression linéaire présentés sur la Figure 49, le modèle de régression linéaire estimé par Epi Info reliant la concentration en ALAT au taux de cholestérol des chiens, sous forme d'une variable en trois classes, s'écrit de la façon suivante :

$$\overline{ALAT}_{/CHOLES\_3CL} = 49{,}696 + 8{,}198.CHOLES\_3CL$$

Cela signifie qu'à partir des données de l'échantillon, le modèle estime que la différence de moyennes de la concentration en ALAT entre deux groupes d'animaux différant de +1 unité pour la variable CHOLES_3CL est de 8,198 UI/L. Ainsi, le modèle estime que la différence entre la moyenne de la concentration en ALAT chez des chiens pour lesquels CHOLES_3CL = 2 (⇔ chiens avec une hypercholestérolémie) et celle chez des chiens pour lesquels CHOLES_3CL = 1 (⇔ chiens avec une normocholestérolémie) est de 8,198 UI/L. Ce modèle estime de la même façon que la différence entre la moyenne de la concentration en ALAT chez des chiens pour lesquels CHOLES_3CL = 1 (⇔ chiens avec une normocholestérolémie) et celle chez des chiens pour lesquels CHOLES_3CL = 0 (⇔ chiens avec une hypocholestérolémie) est *là encore* de 8,198 UI/L.

De la même façon que pour une variable quantitative, inclure une variable qualitative ordinale fait l'hypothèse de la linéarité de l'association entre le CdJ et cette variable qualitative ordinale. Comme pour une variable quantitative, il faudra nécessairement vérifier cette hypothèse avant d'inclure une variable qualitative ordinale dans un modèle de régression (linéaire).

5. *Modèle de régression linéaire univarié avec variable qualitative nominale*

a) *Problématique*

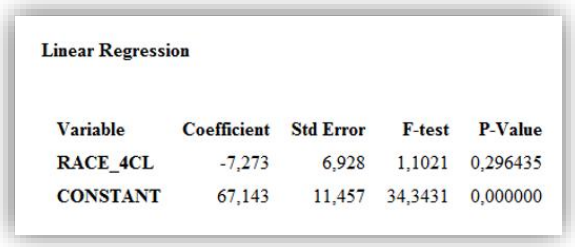
Supposons le modèle de régression linéaire suivant, incluant une seule variable qualitative nominale E :

$$\bar{Y}_{/E} = \alpha + \beta.E$$

Nous allons voir ci-dessous que ce modèle est incorrect, et qu'il fournit une estimation du coefficient β systématiquement ininterprétable. Nous allons illustrer cela à partir du modèle suivant que nous allons faire tourner à partir des données de l'échantillon :

$$\overline{ALAT}_{/RACE\_4CL} = \alpha + \beta.RACE\_4CL$$

Après avoir sélectionné la variable Y du modèle, ici la variable ALAT, et la variable RACE_4CL, nous obtenons les résultats présentés sur la Figure 49.



Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
RACE_4CL	-7,273	6,928	1,1021	0,296435
CONSTANT	67,143	11,457	34,3431	0,000000

Figure 50

A partir des résultats de la régression linéaire présentés sur la Figure 50, le modèle de régression linéaire estimé par Epi Info reliant la concentration en ALAT à la race des chiens, sous forme d'une variable en quatre classes, s'écrit de la façon suivante :

$$\overline{ALAT}_{/RACE\_4CL} = 67{,}143 - 7{,}273.RACE\_4CL$$

Cela signifie qu'à partir des données de l'échantillon, le modèle estime que la différence de moyennes de la concentration en ALAT entre deux groupes d'animaux différant de +1 unité pour la variable RACE_4CL est de -7,273 UI/L. Ainsi, le modèle estime que la différence entre la moyenne de la concentration en ALAT chez des chiens pour lesquels RACE_4CL = 3 (⇔ autre race) et celle chez des chiens pour lesquels RACE_4CL = 2 (⇔ race croisée Golden/Labrador), que celle entre la moyenne de la concentration en ALAT chez des chiens pour lesquels RACE_4CL = 2 (⇔ race croisée Golden/Labrador) et celle chez des chiens pour lesquels RACE_4CL = 1 (⇔ race Labrador), et que celle entre la moyenne de la concentration en ALAT chez des chiens pour lesquels RACE_4CL = 1 (⇔ race Labrador) et celle chez des chiens pour lesquels RACE_4CL = 0 (⇔ race Golden), valent *toutes* -7,273 UI/L. Cette estimation de -7,273 UI/L n'a aucun sens, car elle repose sur l'hypothèse, n'ayant *aucun* fondement physiologique, d'une *même* différence de moyennes de concentration en ALAT lorsque la variable RACE_4CL augmente de +1 unité.

Il est par conséquent *interdit* d'inclure une variable qualitative nominale dans un modèle de régression (linéaire), car ce modèle part du principe que l'augmentation de +1 unité de cette variable a un sens, alors qu'il n'en a aucun dans le cas d'une variable qualitative *nominale*. En effet, le choix du chiffre attribué à une classe d'une variable qualitative nominale est purement arbitraire, et n'est en aucun cas fondé sur un ordre quelconque (« 1 », plus petit que « 2 », lui-même plus petit que « 3 », etc).

b) Création de « dummy variables » et interprétation théorique

Pour étudier l'association entre un CdJ et une variable qualitative nominale à l'aide d'un modèle de régression (quel que soit le modèle de régression), il faut créer autant de variables binaires (« dummy variables » en anglais) que de classes de cette variable qualitative nominale à partir des valeurs de la variable qualitative nominale, puis inclure *K*-1 de ces « dummy variables », avec *K* le nombre de classes de la variable qualitative nominale. Ce que je viens d'écrire peut être dit autrement : parmi les *K* « dummy variables » créées, il faut en exclure une du modèle et inclure toutes les autres. Cette « dummy variable » que l'on *choisit* de ne pas inclure dans le modèle est considérée comme la *classe de référence* de la variable qualitative nominale initiale, et nous allons tout de suite voir pourquoi.

Le tableau ci-dessous présente la façon générale de créer ces « dummy variables », nommées $DUMMY\_VAR_i$, à partir d'une variable qualitative dont la variable correspondante est nommée VAR_QUAL. Ce tableau présente des « 1 » sur la diagonale, et des « 0 » partout ailleurs.

VAR_QUAL	$DUMMY\_VAR_1$	$DUMMY\_VAR_2$	…	$DUMMY\_VAR_i$	…	$DUMMY\_VAR_K$
1	1	0	…	0	…	0
2	0	1	…	0	…	0
…	…	…	…	…	…	…
i	0	0	…	1	…	0
…	…	…	…	…	…	…
K	0	0	…	0	…	1

Reprenons l'exemple de la variable RACE_4CL. Dans la mesure où cette variable comprend quatre classes, il faut créer quatre « dummy variables », que nous allons nommer : GOLDEN, LABRADOR, CROISEE, AUTRE_RACE. L'attribution des valeurs « 0 » et « 1 » pour chacune de ces quatre variables est décrite ci-dessous.

RACE_4CL	GOLDEN	LABRADOR	CROISEE	AUTRE_RACE
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1

Ainsi, un chien de race Labrador se verra attribuer les valeurs de « 0 », « 1 », « 0 », et « 0 », respectivement pour les variables GOLDEN, LABRADOR, CROISEE, et AUTRE_RACE. Ces quatre variables peuvent tout à fait être créées dans le fichier Excel, avant d'être importé dans Epi Info. Nous allons voir aussi qu'Epi Info est capable de créer, lui-même, ces « dummy variables ».

J'ai écrit ci-dessus qu'il fallait ensuite inclure *K*-1 « dummy variables » parmi les *K* créées. Dans l'exemple, nous devons inclure trois parmi les quatre variables créées (GOLDEN, LABRADOR, CROISEE, et AUTRE_RACE). Comme je l'ai écrit, le choix de la « dummy variable » qui ne sera pas incluse dans le modèle vous revient totalement. Supposons que l'on choisisse de ne pas inclure la variable GOLDEN dans le modèle. Le modèle est donc celui-ci :

$$\overline{ALAT}_{/LABRADOR,CROISEE,AUTRE\_RACE} = \alpha + \beta_{LAB}.LABRADOR + \beta_{CROISEE}.CROISEE + \beta_{AUTRE\_R}.AUTRE\_RACE$$

Pour interpréter chacun des trois coefficients du modèle (β_{LAB}, $\beta_{CROISEE}$, β_{AUTRE_R},), je vais écrire le modèle pour chaque race de chien.

Pour les chiens de race Golden :

$$\overline{ALAT}_{/LABRADOR=0,CROISEE=0,AUTRE\_RACE=0} = \alpha + \beta_{LAB} \times 0 + \beta_{CROISEE} \times 0 + \beta_{AUTRE\_R} \times 0 = \alpha$$

Pour les chiens de race Labrador :

$$\overline{ALAT}_{LABRADOR=1,CROISEE=0,AUTRE\_RACE=0} = \alpha + \beta_{LAB} \times 1 + \beta_{CROISEE} \times 0 + \beta_{AUTRE\_R} \times 0 = \alpha + \beta_{LAB}$$

Pour les chiens de race croisée Golden/Labrador :

$$\overline{ALAT}_{LABRADOR=0,CROISEE=1,AUTRE\_RACE=0} = \alpha + \beta_{LAB} \times 0 + \beta_{CROISEE} \times 1 + \beta_{AUTRE\_R} \times 0 = \alpha + \beta_{CROISEE}$$

Pour les chiens d'autre race :

$$\overline{ALAT}_{LABRADOR=0,CROISEE=0,AUTRE\_RACE=1} = \alpha + \beta_{LAB} \times 0 + \beta_{CROISEE} \times 0 + \beta_{AUTRE\_R} \times 1 = \alpha + \beta_{AUTRE\_R}$$

Comme précédemment, je vais faire la soustraction de deux modèles pour interpréter chaque coefficient β.

$$\overline{ALAT}_{/LABRADOR=1,CROISEE=0,AUTRE\_RACE=0} - \overline{ALAT}_{/LABRADOR=0,CROISEE=0,AUTRE\_RACE=0} = \beta_{LAB}$$

$$\overline{ALAT}_{/LABRADOR=0,CROISEE=1,AUTRE\_RACE=0} - \overline{ALAT}_{/LABRADOR=0,CROISEE=0,AUTRE\_RACE=0} = \beta_{CROISEE}$$

$$\overline{ALAT}_{/LABRADOR=0,CROISEE=0,AUTRE\_RACE=1} - \overline{ALAT}_{/LABRADOR=0,CROISEE=0,AUTRE\_RACE=0} = \beta_{AUTRE\_R}$$

Ainsi, β_{LAB} est la valeur de la différence de moyennes de concentration en ALAT entre les chiens de race Labrador et les chiens de race Golden, $\beta_{CROISEE}$ est la valeur de la différence de moyennes de concentration en ALAT entre les chiens de race croisée et les chiens de race Labrador, et β_{AUTRE_R} est la valeur de la différence de moyennes de concentration en ALAT entre les chiens d'autre race et les chiens de race Labrador. Vous venez de voir que chacun des coefficients du modèle compare une classe de la variable RACE_4CL à la *classe de référence* qui correspond à la « dummy variable » qui n'a pas été incluse dans le modèle, à savoir ici la classe « Golden ». Si nous avions choisi de ne pas inclure la « dummy variable » LABRADOR, alors chacun des trois coefficients du modèle aurait quantifié la différence de moyennes de concentration en ALAT entre chacune des classes de la variable RACE_4CL et la classe de référence correspondant à la race Labrador.

c) *Mise en pratique avec Epi Info*

Nous allons voir désormais comment inclure des « dummy variables » dans un modèle de régression (linéaire), pour étudier l'association entre un CdJ (ici quantitatif, la concentration en ALAT) et une variable qualitative nominale (ici, la race des chiens, en 4 classes), avec Epi Info.

Pour cela (cf. Figure 51), on sélectionne la variable Y du modèle, ici la variable ALAT (a), on sélectionne la variable que l'on veut inclure dans le modèle sous forme de « dummy variables », ici la variable RACE_4CL (b), mais surtout on ne s'arrête pas là, et l'on clique sur « RACE_4CL », ce qui fait apparaître « Make Dummy » (c). Après avoir cliqué sur « Make Dummy », on voit des parenthèses apparaître autour de « RACE_4CL » (d), ce qui indique bien que des « dummy variables » vont être créées par Epi Info au moment de faire tourner le modèle. Puis on clique sur « Ok » (e).

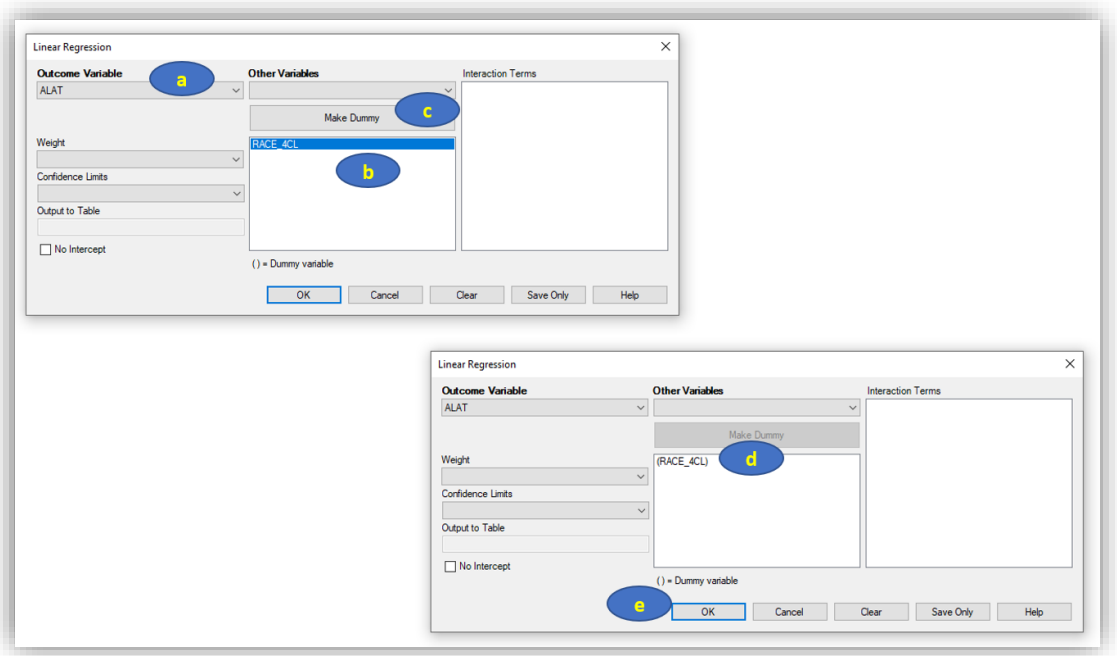


Figure 51

On obtient les résultats présentés sur la Figure 52 ci-dessous.

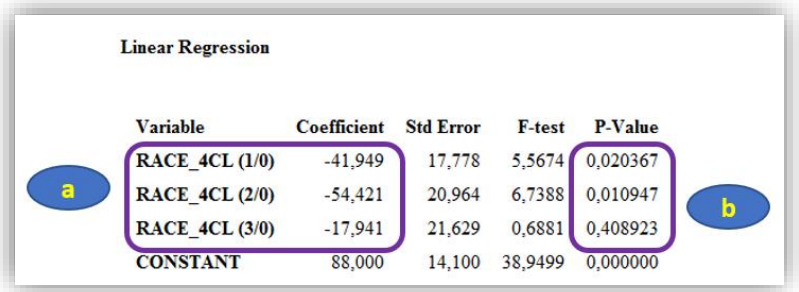


Figure 52

Pour la suite de ce guide, et dans la mesure où je vais encore inclure dans un modèle de régression une variable qualitative (nominale) à l'aide de la commande « Make Dummy » dans Epi Info, voici comment je vais choisir d'écrire le modèle qui correspond à ce que j'ai sélectionné dans la Figure 51, en mettant entre parenthèses la variable qualitative (nominale) que l'on transforme en « dummy variables » :

$$\overline{ALAT}_{/(RACE\_4CL)} = \alpha + \beta_{1,2,3}.(RACE\_4CL)$$

Les résultats présentés sur la Figure 52, qui découlent de la sélection sur la Figure 51, peuvent s'écrire ainsi :

$$\overline{ALAT}_{/(RACE\_4CL)} = \alpha + \beta_1.RACE\_4CL(1/0) + \beta_2.RACE\_4CL(2/0) + \beta_3.RACE\_4CL(3/0)$$

Avec β_1 = -41,949, β_2 = -54,421, et β_3 = -17,941 (cf. Figure 52.a).

Là, il va être fondamental de bien comprendre ces résultats. Car sinon, vous pourriez dire vraiment de belles bêtises à partir des degrés de signification présentés sur la Figure 52. Notamment, il faut comprendre ces « (1/0) », « (2/0) », et « (3/0) » à droite de « RACE_4CL » dans la colonne « Variable » (Figure 52.a).

Dans la mesure où la variable qualitative nominale RACE_4CL a été incluse sous forme de « dummy variables » (et c'est obligatoire de faire cela car il s'agit d'une variable qualitative *nominale*), chaque coefficients β_i du modèle (ici, $i \in \{1,2,3\}$) quantifie la différence de moyennes sur la concentration en ALAT entre une des classes de la variable qualitative et la *classe de référence*. La classe de référence est *choisie* par Epi Info de la façon suivante : c'est la classe dont la valeur du codage est la plus faible. Ici, la classe de référence choisie par Epi Info est donc la classe qui correspond à la valeur « 0 » pour la variable RACE_4CL, soit les chiens de race Golden.

Ainsi, la valeur de β_1 de -41,949 correspond à la différence de moyennes de la concentration en ALAT entre les chiens pour lesquels la variable RACE_4CL vaut « 1 » (⇔ les chiens de race Labrador) et les chiens pour lesquels la variable RACE_4CL vaut « 0 » (⇔ les chiens de race Golden, la classe de référence). Et comme il est indiqué « 1/0 », c'est bien une comparaison des chiens de race Labrador *par rapport* aux chiens de race Golden et non les chiens de race Golden *par rapport* aux chiens de race Labrador. Cette différence étant négative, la moyenne de la concentration en ALAT était moins élevée parmi les chiens de race Labrador que parmi les chiens de race Golden. Cette différence de -41,949 UI/L est d'ailleurs significative : le degré de signification comparant les deux moyennes vaut 0,0204 (Figure 52.b), donc inférieur à 0,05. Par conséquent, on peut aussi interpréter cette valeur de -41,949, significativement différente de « 0 », de la façon suivante : les chiens de race Labrador avaient une moyenne de concentration en ALAT significativement inférieure à celle des chiens de race Golden (Δ = -41,949 UI/L ; p = 0,02).

De même, la valeur de la valeur de β_2 de -54,421 correspond à la différence de moyennes de la concentration en ALAT entre les chiens pour lesquels la variable RACE_4CL vaut « 2 » (⇔ les chiens de race croisée Golden/Labrador) et les chiens pour lesquels la variable RACE_4CL vaut « 0 » (⇔ les chiens de race Golden, la classe de référence). Cette différence étant là encore négative et significative, la moyenne de la concentration en ALAT était significativement moins élevée parmi les chiens de race croisée Golden/Labrador que parmi les chiens de race Golden.

Enfin, la valeur de la valeur de β_3 de -17,941 correspond à la différence de moyennes de la concentration en ALAT entre les chiens pour lesquels la variable RACE_4CL vaut « 3 » (les chiens d'autres races) et les chiens pour lesquels la variable RACE_4CL vaut « 0 » (les chiens de race Golden, la classe de référence). Cette différence étant là encore négative (mais cette fois-ci non significative), la moyenne de la concentration en ALAT était moins élevée parmi les chiens d'autres races que parmi les chiens de race Golden.

Vous vous rendez compte que la pertinence de l'interprétation de ces résultats dépend totalement du choix de la classe de référence (choisie par Epi Info à partir du codage de la variable qualitative nominale que, pour le coup, *vous* avez choisi). Par conséquent, la façon de choisir *vous-même* votre classe de référence en utilisant Epi Info et l'option « Make Dummy » (Figure 51.c) est de modifier le codage de la variable qualitative nominale dans le fichier Excel *avant* d'être importé dans Epi Info. Par exemple, si on avait voulu que ce soit les chiens de race croisée qui soient considérés comme « classe de référence », il aurait fallu créer la variable RACE_4CL dont la plus petite valeur (par exemple « 0 ») aurait été attribuée à ces chiens là (les chiffres « 1 », « 2 », et « 3 » étant attribués aux classes restantes, selon votre choix). Une autre façon de faire, c'est de créer vous-mêmes les « dummy

variables » dans Excel avant l'import dans Epi Info, et de choisir *vous-même* d'inclure les *K*-1 « dummy variables ». C'est ce que je présente ci-dessous, à partir des quatre variables LABRADOR, GOLDEN, CROISEE, et AUTRE_RACE qui se trouvent dans le fichier de données (cf. Chapitre 1, Partie III, page 9).

Pour retrouver les résultats présentés sur la Figure 52, nous allons inclure les trois « dummy variables » que sont GOLDEN, CROISEE, et AUTRE_RACE, en sélectionnant ces trois variables, et donc choisir de ne pas inclure la variable LABRADOR dans le modèle (cf. Figure 53).

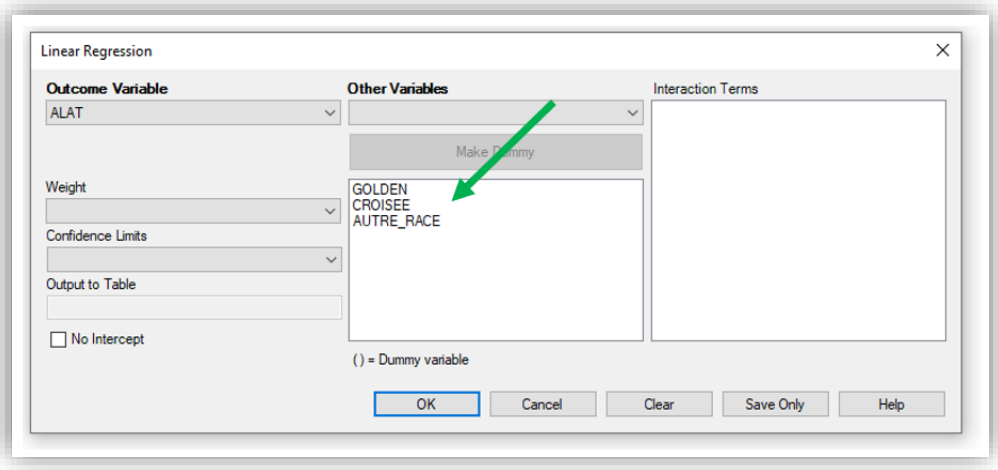
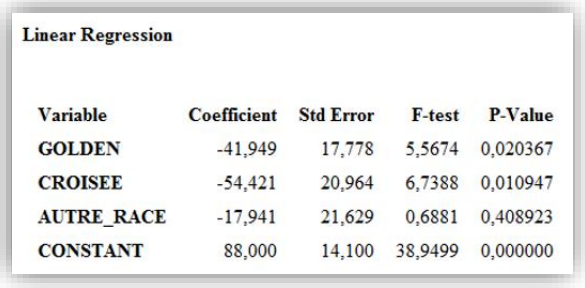


Figure 53

Les résultats de cette régression linéaire sont présentés sur la Figure 54 ci-dessous.



Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
GOLDEN	-41,949	17,778	5,5674	0,020367
CROISEE	-54,421	20,964	6,7388	0,010947
AUTRE_RACE	-17,941	21,629	0,6881	0,408923
CONSTANT	88,000	14,100	38,9499	0,000000

Figure 54

On retrouve évidemment les mêmes résultats que ceux présentés sur la Figure 52 : le coefficient β devant « GOLDEN » sur la Figure 54 est égal au coefficient β devant « RACE_4CL(1/0) » sur la Figure 52 (et idem pour les deux autres coefficients β).

C. Interprétation des résultats d'une régression linéaire multivariée

1. *Interprétation générale*

Si, maintenant, le modèle inclut deux variables E_1 et E_2, il devient alors :

$$\bar{Y}_{/E_1,E_2} = \alpha + \beta_1.E_1 + \beta_2.E_2$$

L'interprétation de β_1 devient : « β_1 est la valeur, estimée à partir des données de l'échantillon, de la différence de moyennes de Y entre deux groupes d'animaux différant de +1 unité pour leur variable E_1, ajustée sur E_2. » L'expression « ajustée sur E_2 » est équivalente à « indépendamment de E_2 » ou « après avoir pris en compte E_2 ». Je ne vais pas faire, dans ce guide, la démonstration qui prouve que

le fait d'inclure la variable E_2 dans le modèle conduit à ce que β_1 quantifie l'association entre E_1 et Y *ajustée* sur E_2. Vous devez me faire confiance 😊…

Si, enfin, le modèle inclut N variables E_1, E_2 , … , E_N, il devient alors :

$$\bar{Y}_{/E_1,E_2,\ldots,E_N} = \alpha + \sum_{i=1}^{N} \beta_i . E_i$$

L'interprétation de β_i devient : « β_i est la valeur, estimée à partir des données de l'échantillon, de la différence de moyennes de Y entre deux groupes d'animaux différant de +1 unité pour leur variable E_i, ajustée sur toutes les autres variables incluses dans le modèle. »

L'ordre des variables incluses dans un modèle multivarié n'a aucune importance. Ainsi, que l'on souhaite étudier l'association entre Y et E_1, ajustée sur E_2 et E_3, ou bien que l'on souhaite étudier l'association entre Y et E_2, ajustée sur E_1 et E_3, dans les deux cas le modèle de régression (linéaire) sera identique :

$$\bar{Y}_{/E_1,E_2,E_3} = \alpha + \beta_1 . E_1 + \beta_2 . E_2 + \beta_3 . E_3$$

2. *En pratique avec Epi Info*

Supposons que l'on veuille étudier l'association entre la concentration en ALAT et le sexe des chiens, ajustée sur l'âge et la race des chiens. Le modèle de régression linéaire multivarié correspondant est donc le suivant :

$$\overline{ALAT}_{/FEMELLE,AGE,(RACE\_4CL)} = \alpha + \beta . FEMELLE + \gamma . AGE + \delta_{1,2,3} . (RACE\_4CL)$$

Je rappelle vivement deux choses : dans ce modèle, la variable RACE_4CL doit obligatoirement être incluse dans le modèle sous forme de « dummy variables » (d'où les parenthèses autour de « RACE_4CL » ci-dessus), et ensuite, tous les coefficients du modèle (β, γ, δ_1, δ_2, et δ_3) ne sont interprétables que si le modèle repose sur des hypothèses vérifiées : (1) la concentration en ALAT doit suivre une loi normale, et (2) l'association entre l'âge et la concentration en ALAT doit être linéaire.

Pour faire tourner le modèle ci-dessus, voici ce qu'il faut sélectionner dans Epi Info (cf. Figure 55) :

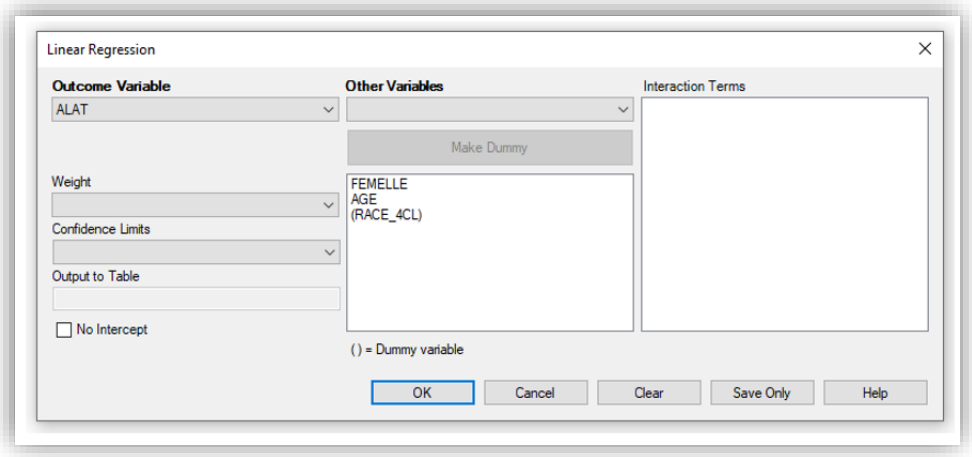


Figure 55

On obtient les résultats présentés sur la Figure 56.

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
FEMELLE	12,169	13,860	0,7710	0,382210
AGE	1,531	2,538	0,3640	0,547787
RACE_4CL (1/0)	-41,717	18,095	5,3153	0,023387
RACE_4CL (2/0)	-56,044	21,590	6,7383	0,010984
RACE_4CL (3/0)	-20,962	22,190	0,8924	0,347290
CONSTANT	69,272	25,084	7,6267	0,006943

Figure 56

A partir des résultats de la régression linéaire présentés sur la Figure 56, le modèle de régression linéaire estimé par Epi Info reliant la concentration en ALAT au sexe, à l'âge, et à la race, s'écrit de la façon suivante (à partir de la colonne « Coefficient » sur la Figure 56) :

$$\overline{ALAT}_{/FEMELLE,AGE,(RACE\_4CL)} = 69{,}272 + 12{,}169.FEMELLE + 1{,}531.AGE - 41{,}717.RACE\_4CL(1/0) - 56{,}044.RACE\_4CL(2/0) - 20{,}962.RACE\_4CL(3/0)$$

J'insiste sur le fait que vous devez impérativement avoir conscience que dès que l'on inclut dans un modèle une variable qualitative ordinale ou quantitative telle quelle (c'est-à-dire, sans l'inclure sous forme de « dummy variables »), *tous* les coefficients du modèle (y compris ceux associés à des variables binaires) ne sont interprétables que si l'hypothèse de la linéarité de l'association pour chacune des variables qualitatives ordinales ou quantitatives incluses dans le modèle est vérifiée.

Nous allons faire l'hypothèse que l'association entre l'âge et la concentration en ALAT est linéaire. Interprétons, maintenant que nous pouvons le faire, chacun des coefficients associés aux variables incluses dans le modèle multivarié. La valeur de 12,169 devant la variable FEMELLE signifie qu'indépendamment de l'âge et de la race, dans l'échantillon, les chiens femelles avaient une concentration en ALAT moyenne supérieure de 12,169 UI/L à celle des chiens mâles. La valeur de 1,531 devant la variable AGE signifie qu'indépendamment du sexe et de la race, dans l'échantillon, une augmentation de +1 année d'âge est associée à une augmentation moyenne en ALAT de 1,531 UI/L. La valeur de -41,717 devant la « dummy variable » RACE_4CL(1/0) signifie qu'indépendamment du sexe et de l'âge, dans l'échantillon, les chiens de race Labrador (classe « 1 » pour RACE_4CL) avaient une concentration en ALAT moyenne *inférieure* de 41,717 UI/L à celle des chiens de race Golden (classe « 0 » pour RACE_4CL, qui est la classe de référence). La valeur de -56,044 devant la « dummy variable » RACE_4CL(2/0) signifie qu'indépendamment du sexe et de l'âge, dans l'échantillon, les chiens de race croisée Golden/Labrador (classe « 2 » pour RACE_4CL) avaient une concentration en ALAT moyenne *inférieure* de 56,044 UI/L à celle des chiens de race Golden. La valeur de -20,962 devant la « dummy variable » RACE_4CL(3/0) signifie qu'indépendamment du sexe et de l'âge, dans l'échantillon, les chiens d'autre race (classe « 3 » pour RACE_4CL) avaient une concentration en ALAT moyenne *inférieure* de 20,962 UI/L à celle des chiens de race Golden.

IV. Vérification de la linéarité de l'association avec une variable quantitative ou qualitative ordinale

A. Introduction

La première question à se poser avant de vérifier la linéarité de l'association avec le CdJ (quel qu'en soit le type : quantitatif ou binaire, assorti ou non d'un temps de survie) est de savoir s'il y a des raisons physiologiques ou physiopathologiques qui pourraient conduire à une association réellement non linéaire entre le CdJ et cette variable. Et il est fondamental de tenter de répondre à cette question *avant* la vérification de la linéarité de l'association. En effet, il faut garder en tête que ce que l'on peut observer dans un échantillon peut être éloigné de ce qu'il se passe dans la population cible, entre autres à cause de la fluctuation d'échantillonnage. Par conséquent, il ne faut pas *trop* attendre des données qu'elles nous disent si l'association est ou n'est pas linéaire – il faut que nous en ayons *a priori* une idée.

La démarche de vérification de la linéarité de l'association entre l'état de santé Y et une variable qualitative ordinale ou quantitative peut être considérée comme fastidieuse. Elle est néanmoins indispensable à réaliser si l'on souhaite au final inclure dans un modèle de régression (quel qu'il soit) une telle variable. Inclure une variable qualitative ordinale ou quantitative dans un modèle et interpréter les résultats de ce modèle sans avoir vérifié au préalable cette hypothèse de la linéarité de l'association expose l'auteur.e de ce modèle à des interprétations fausses, et par conséquent à des erreurs de communication scientifique.

Si, après avoir lu ce qui suit, vous trouvez effectivement que la démarche est *trop* fastidieuse, alors vous n'avez plus qu'une seule solution : utiliser des variables uniquement binaires. Cela signifie que si vos variables d'intérêt et/ou vos facteurs de confusion ne sont pas des variables binaires, vous *devrez* les recoder en variables binaires, soit selon un seuil déjà décrit dans la littérature, ou bien selon un seuil qui a cliniquement du sens, ou enfin selon la médiane ou selon le premier ou troisième quartile. L'inconvénient de rendre binaire une variable initialement qualitative ordinale ou quantitative est entre autres décrit dans les articles suivants (Altman and Royston, 2006; Brenner and Blettner, 1997; Royston et al., 2006).

Je vais présenter ci-dessous une démarche pour vérifier l'hypothèse de la linéarité de l'association, dans le cas de la régression linéaire. Mais cette démarche est absolument identique quel que soit le modèle de régression.

B. Cas d'une variable qualitative ordinale

1. Aspect théorique

Soit VAR_QUAL_K_CL une variable qualitative ordinale à *K* classes, codée « 0 », « 1 », …, « K-1 ». Pour vérifier que l'association entre le CdJ quantitatif (quantifié par Y) et la variable VAR_QUAL_K_CL est linéaire (c'est-à-dire, qu'une augmentation de +1 unité de VAR_QUAL_K_CL se traduit par une même augmentation sur Y, quelle que soit la valeur de VAR_QUAL_K_CL), je vous recommande de suivre la démarche suivante :

1) Créer les « dummy variables » à partir de la variable qualitative ordinale dont on souhaite montrer la linéarité de l'association avec l CdJ (Epi Info le fait tout seul, en cliquant sur « Make Dummy »).

2) Inclure *K*-1 « dummy variables » dans le modèle (avec *K* le nombre de classes de la variable qualitative ordinale), et notamment la $2^{ème}$, $3^{ème}$, …, $K^{ème}$ (je recommande en effet que ce soit la première « dummy variable » qui ne soit pas incluse dans le modèle, et qui soit donc considérée comme la « classe de référence » ; et cela tombe bien, puisque c'est ce que fait Epi Info tout seul).

3) Noter les valeurs des coefficients β (ainsi que la SE_β) de chacune des *K*-1 « dummy variables ».

4) Placer sur un graphique *K*-1 points, chacun ayant pour ordonnée la valeur du coefficient β d'une « dummy variable » et pour abscisse la valeur représentant la classe de variable qualitative ordinale, puis placez autour de chacun de ces points l'$IC_{95\%}$ de β.

5) Placer le point d'ordonnée 0 et d'abscisse la valeur représentant la 1ère classe de variable qualitative ordinale (la classe de référence).

6) Vérifier visuellement que tous les points du graphique (incluant le point d'ordonnée 0) sont relativement bien alignés sur une droite.

Si l'association entre le CdJ (quantifié par Y) et VAR_QUAL_K_CL est *parfaitement* linéaire (mathématiquement parlant), la Figure 57 présente ce que l'on devrait obtenir comme valeurs des *K*-1 coefficients βi (i ∈ {1, …, *K*-1}) : ils devraient être tous alignés sur une droite de pente β (en passant par le point d'abscisse la valeur de la 1ère classe, ici 0, et d'ordonnée 0, comme l'étape 5 ci-dessus le demande).

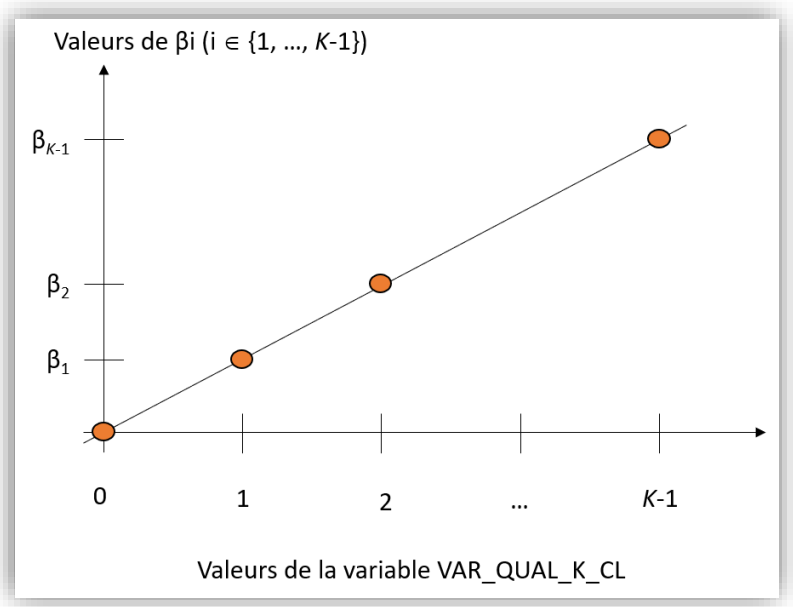


Figure 57

Je vais passer un peu de temps sur l'étape 6 ci-dessus. Tout d'abord, vous pouvez tout à fait vous aider des $IC_{95\%}$ des coefficients β qui donnent une indication de la précision avec laquelle ces coefficients β ont été estimés. En effet, si en vrai l'association est linéaire, la fluctuation d'échantillonnage peut conduire à des estimations des coefficients β éloignées des valeurs réelles, et ce d'autant plus que chaque classe de la variable qualitative nominale est composée de peu d'individus. Ainsi, les coefficients β peuvent ne pas être alignés non pas parce que l'association n'est réellement pas linéaire, mais tout simplement à cause d'une imprécision des estimations des coefficients β à partir des données de l'échantillon. La Figure 58 présente deux situations, avec des valeurs des coefficients β identiques, mais avec des SE_β (donc des $IC_{95\%}$, représentés sur la figure) différents.

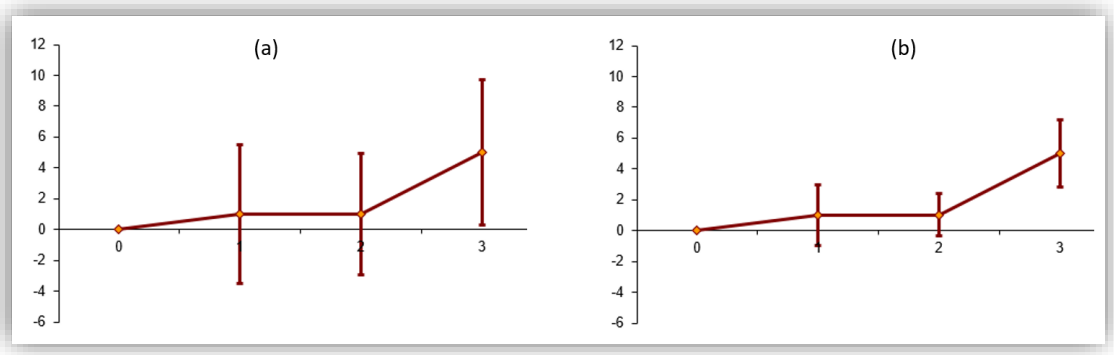


Figure 58

La situation (a) est celle où l'on pourrait accepter l'hypothèse de la linéarité de l'association dans le cas de figure où l'on n'aurait aucune raison de penser qu'en vrai, l'association n'est pas linéaire. En effet, bien que les quatre points ne soient pas vraiment alignés, la largeur des $IC_{95\%}$ laisse penser que ce non alignement semble davantage dû à une imprécision des 3 estimations des coefficients qu'à une non linéarité réelle. La situation (b) laisse en revanche penser que l'association n'est pas linéaire.

Si l'hypothèse de la linéarité de l'association est vérifiée pour VAR_QUAL_K_CL, alors il est possible de faire tourner le modèle de régression linéaire incluant cette variable telle quelle : le coefficient β associé à VAR_QUAL_K_CL sera interprétable (augmentation moyenne de Y pour une augmentation de +1 unité de VAR_QUAL_K_CL).

Si l'hypothèse de la linéarité de l'association n'est pas vérifiée, alors il faut soit inclure la variable VAR_QUAL_K_CL sous forme de « dummy variables », soit rendre la VAR_QUAL_K_CL de façon binaire, en regroupant les classes entre elles. Ce regroupement doit d'abord être guidé par la « clinique ». Ce regroupement doit en effet et avant tout être cliniquement pertinent. S'il est guidé par la représentation graphique des points, c'est dangereux. En effet, recoder des variables doit se faire *a priori*, et non pas *après* avoir vu les résultats. La démonstration serait trop longue ici, et sort de toute façon du cadre de ce guide.

Sur la Figure 57, nous avons utilisé comme abscisses des points les valeurs des classes de VAR_QUAL_K_CL. Dans le cas où VAR_QUAL_K_CL a été codée à partir d'une variable quantitative (comme c'est le cas pour la variable UREE_4CL du fichier de données), nous verrons qu'au lieu d'utiliser les valeurs des classes de VAR_QUAL_K_CL, je vous recommande d'utiliser le « centre » des classes de VAR_QUAL_K_CL. Ce « centre » d'une classe i de VAR_QUAL_K_CL pourra être la médiane de la variable quantitative (dont est issue VAR_QUAL_K_CL) calculée parmi tous les individus pour lesquels VAR_QUAL_K_CL = i.

2. *En pratique avec Epi Info*

Prenons l'exemple de l'association entre la concentration en ALAT et celle en cholestérol en trois classes (hypocholestérolémie, normocholestérolémie, et hypercholestérolémie), à l'aide de la variable CHOLES_3CL. En incluant la variable CHOLES_3CL dans le modèle, celui-ci s'écrit ainsi :

$$\overline{ALAT}_{/CHOLES\_3CL} = \alpha + \beta.CHOLES\_3CL$$

Comme nous l'avons vu ci-dessus (cf. Figure 49), l'estimation de β de valeur égale à 8,198 n'a de sens que si l'association entre la concentration en ALAT et la variable CHOLES_3CL est physiopathologiquement linéaire (une augmentation de +1 unité pour la variable CHOLES_3CL se traduit par une même augmentation de la concentration en ALAT). Nous allons désormais vérifier cette hypothèse de linéarité de l'association.

Faisons tourner le modèle suivant dans Epi Info :

$$\overline{ALAT}_{/(CHOLES\_3CL)} = \alpha + \beta_{1,2}.(CHOLES\_3CL)$$

Dans Epi Info, ce modèle ci-dessus correspond à celui-ci-dessous :

$$\overline{ALAT}_{/(CHOLES\_3CL)} = \alpha + \beta_{1}.CHOLES\_3CL(1/0) + \beta_{2}.CHOLES\_3CL(2/0)$$

Pour cela, après avoir sélectionné la variable CHOLES_3CL dans Epi Info, il faut cliquer sur « Make Dummy » (cf. Figure 59).

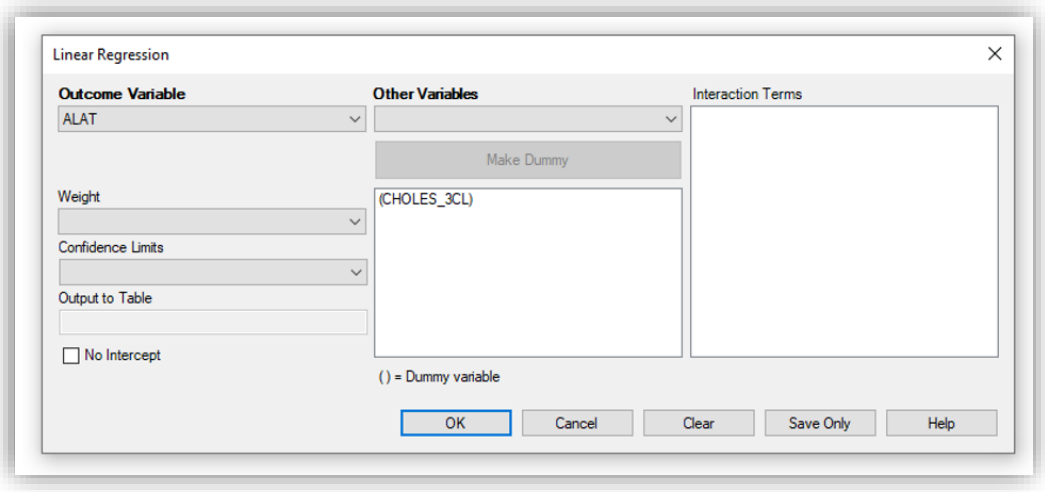
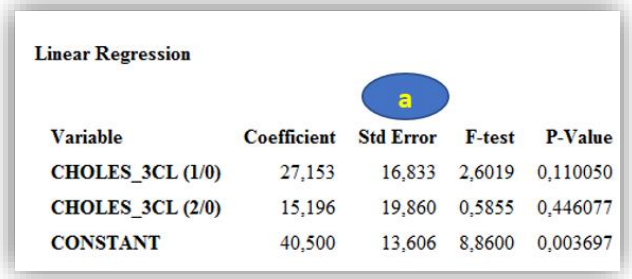


Figure 59

La Figure 60 présente les résultats que l'on obtient.



Linear Regression

a

Variable	Coefficient	Std Error	F-test	P-Value
CHOLES_3CL (1/0)	27,153	16,833	2,6019	0,110050
CHOLES_3CL (2/0)	15,196	19,860	0,5855	0,446077
CONSTANT	40,500	13,606	8,8600	0,003697

Figure 60

Le modèle estimé s'écrit par conséquent :

$$\overline{ALAT}_{/(CHOLES\_3CL)} = 40{,}500 + 27{,}153.CHOLES\_3CL(1/0) + 15{,}196.CHOLES\_3CL(2/0)$$

Pour vérifier la linéarité de l'association entre la concentration en ALAT et la variable CHOLES_3CL, il faut dresser le graphique de la Figure 57 en plaçant trois points (car CHOLES_3CL a trois classes) : le point d'ordonnée 0, le point d'ordonnée β_1, et le point d'ordonnée β_2. Les abscisses respectives sont les valeurs des trois classes de la variable CHOLES_3CL, à savoir 0, 1, et 2. De plus, comme indiqué dans la partie théorique ci-dessus, je vous recommande de placer les $IC_{95\%}$ des coefficients β_1 et β_2 (en utilisant la SE des coefficients, cf. Figure 60.a). Pour dresser un tel graphique, j'utilise un fichier Excel que je peux vous envoyer si vous m'envoyez une demande à : loic.desquilbet(at)gmail.com.

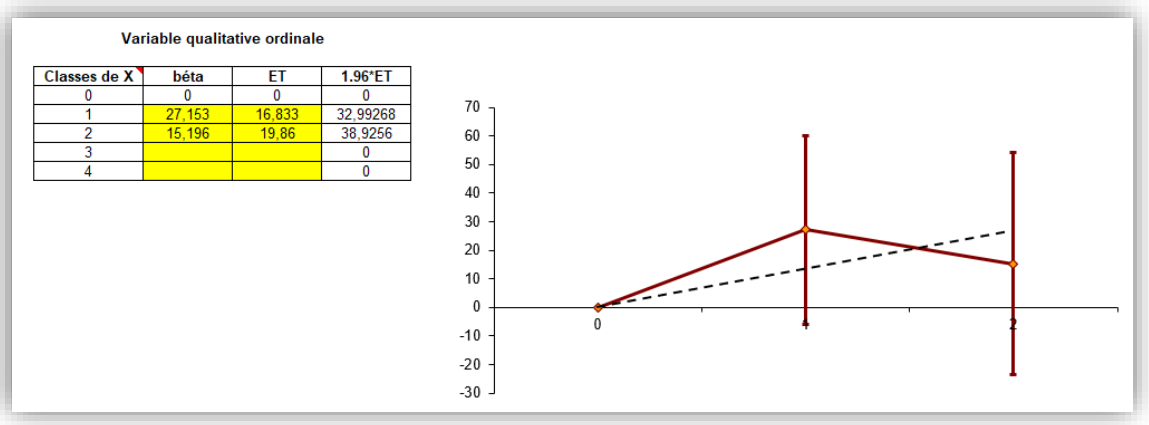


Figure 61

Si l'on ne regarde que l'alignement des trois points, nous dirions que l'association avec la variable CHOLES_3CL n'est pas linéaire. Avec la largeur des $IC_{95\%}$ des deux coefficients β, nous pourrions faire passer une droite passant par le point d'abscisse et d'ordonnée 0 (le point correspondant à la classe de référence, correspondant à CHOLES_3CL=0) et au milieu des deux autres points (trait en pointillé sur la Figure 61). S'il n'y a pas de raison de penser que l'association entre ALAT et CHOLES_3CL n'est *a priori* pas linéaire, alors nous pourrions accepter l'hypothèse de la linéarité de l'association entre ALAT et CHOLES_3CL. Ainsi, sous cette hypothèse, il serait possible de faire tourner le modèle $\overline{ALAT}_{/CHOLES\_3CL} = \alpha + \beta.CHOLES\_3CL$, dont les résultats sont présentés sur la Figure 49 (avec l'interprétation qui se trouve sous cette Figure 49).

Si en revanche il y a des raisons de penser que l'association entre la concentration en ALAT et CHOLES_3CL n'est pas linéaire, physiopathologiquement parlant, alors il faut traiter cette variable qualitative ordinale comme si c'était une variable qualitative nominale, c'est-à-dire faire tourner le modèle avec les « dummy variables », et interpréter les coefficients (ceux de la Figure 60) exactement comme on le ferait avec une variable qualitative nominale. Ou bien il faut étudier l'association avec la cholestérolémie sous format binaire, et non pas en trois classes comme actuellement.

C. Cas d'une variable quantitative

1. Aspect théorique

Pour vérifier la linéarité de l'association avec une variable quantitative, deux étapes sont nécessaires : (1) créer une variable qualitative ordinale à partir de la variable quantitative, et (2) vérifier la linéarité de l'association entre le CdJ et cette variable qualitative ordinale créée.

Si la linéarité de l'association avec la variable qualitative ordinale est vérifiée, alors on fera l'hypothèse que la linéarité de l'association est aussi vérifiée pour la variable quantitative en question. La linéarité de l'association avec une variable quantitative peut être montrée directement sur la variable quantitative, sans passer par la création de la variable qualitative ordinale, mais cela demande plus de travail (Desquilbet and Mariotti, 2010) !

Si la linéarité de l'association avec la variable qualitative ordinale n'est pas vérifiée, alors on fera l'hypothèse que cette linéarité de l'association n'est pas non plus vérifiée pour la variable quantitative.

Pour vérifier la linéarité de l'association avec une variable quantitative, je vous recommande de créer la variable qualitative ordinale à partir des quartiles de la variable quantitative :

Valeur de la variable quantitative	Codage de la variable qualitative ordinale à créer
≤ 1er quartile	0
]1er quartile ; médiane]	1
]Médiane ; 3ème quartile]	2
> 3ème quartile	3

Ensuite, pour vérifier que les coefficients β de chacune des « dummy variables » issues de la variable qualitative ordinale sont alignés, je vous recommande fortement d'attribuer la valeur de la médiane de la variable quantitative pour chaque classe de la variable qualitative ordinale créée pour l'occasion, sur l'axe des abscisses du graphique de vérification de la linéarité de l'association.

2. *En pratique avec Epi Info*

Supposons que l'on veuille étudier l'association entre la concentration en urée et la concentration en ALAT. Je vous rappelle que le modèle $\overline{ALAT}_{/UREE} = \alpha + \beta.UREE$ ne fournit une valeur de β interprétable que si l'association entre UREE et ALAT est linéaire. Nous allons vérifier cela ci-dessous. Pour cela, nous devons créer une variable qualitative en quatre classes selon les quartiles de la variable UREE. Il se trouve que le fichier de données contient déjà cette variable, qui se nomme UREE_4CL (cf. Chapitre 1, Partie III, page 9). Il s'agit donc de vérifier la linéarité de l'association entre la variable ALAT et la variable UREE_4CL. Nous allons donc faire tourner le modèle suivant dans Epi Info :

$$\overline{ALAT}_{/(UREE\_4CL)} = \alpha + \beta_{1,2,3}.(UREE\_4CL)$$

Dans Epi Info, ce modèle ci-dessus correspond à celui-ci-dessous :

$$\overline{ALAT}_{/(UREE\_4CL)} = \alpha + \beta_1.UREE\_4CL(1/0) + \beta_2.UREE\_4CL(2/0) + \beta_3.UREE\_4CL(3/0)$$

Pour cela, après avoir sélectionné la variable UREE_4CL dans Epi Info, il faut cliquer sur « Make Dummy » (cf. Figure 62).

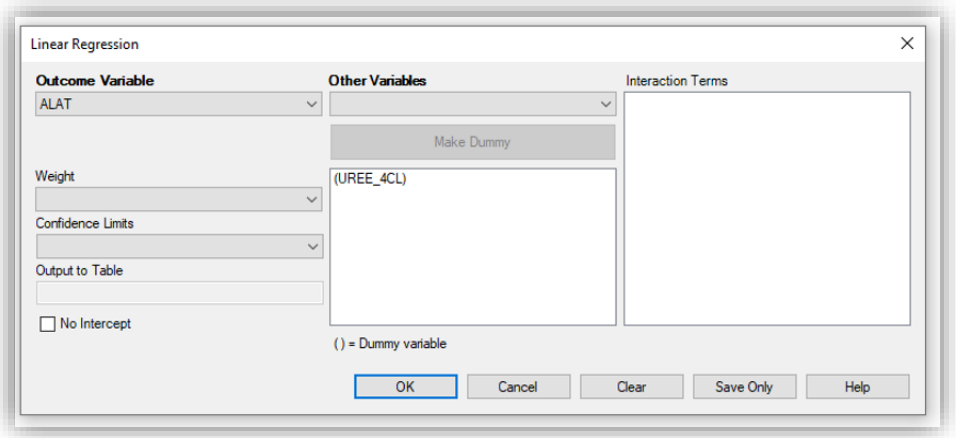


Figure 62

La Figure 63 présente les résultats que l'on obtient.

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
UREE_4CL (1/0)	25,920	20,442	1,6077	0,207940
UREE_4CL (2/0)	9,779	20,054	0,2378	0,626934
UREE_4CL (3/0)	7,007	20,241	0,1198	0,729988
CONSTANT	46,913	14,608	10,3136	0,001808

Figure 63

Pour vérifier la linéarité de l'association entre la concentration en ALAT et la variable UREE_4CL, il faut dresser le graphique de la Figure 57 en plaçant quatre points (car UREE_4CL a quatre classes) : le point d'ordonnée 0, le point d'ordonnée β_1, le point d'ordonnée β_2, et le point d'ordonnée β_3. Mais contrairement à la situation de la vérification de l'association avec une variable qualitative ordinale originale, dans cette situation où la variable qualitative ordinale provient d'une variable quantitative, je vous recommande de placer ces quatre points avec comme abscisse la médiane de la variable quantitative (ici, la médiane de UREE) pour chacune des quatre classes de la variable UREE_4CL. Pour cela, il suffit de suivre ce qui est indiqué dans la Partie III.C.2 du Chapitre 2, page 27 (cf. Figure 64).

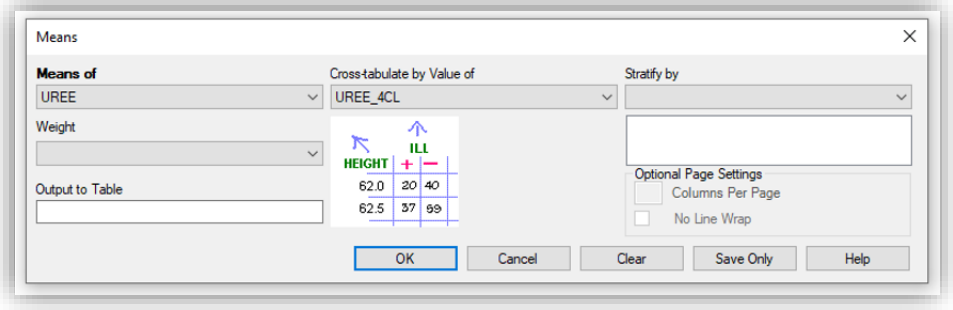
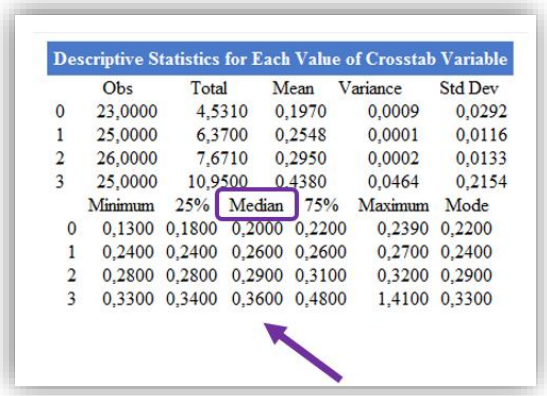


Figure 64

Après avoir cliqué que « Ok », nous obtenons les résultats ci-dessous (Figure 65).



Descriptive Statistics for Each Value of Crosstab Variable

	Obs	Total	Mean	Variance	Std Dev
0	23,0000	4,5310	0,1970	0,0009	0,0292
1	25,0000	6,3700	0,2548	0,0001	0,0116
2	26,0000	7,6710	0,2950	0,0002	0,0133
3	25,0000	10,9500	0,4380	0,0464	0,2154

	Minimum	25%	Median	75%	Maximum	Mode
0	0,1300	0,1800	0,2000	0,2200	0,2390	0,2200
1	0,2400	0,2400	0,2600	0,2600	0,2700	0,2400
2	0,2800	0,2800	0,2900	0,3100	0,3200	0,2900
3	0,3300	0,3400	0,3600	0,4800	1,4100	0,3300

Figure 65

Le centre des classes de la variable UREE_4CL que nous allons donc utiliser pour dresser le graphique de la Figure 57 sont donc de 0,20, 0,26, 0,29, et 0,36 respectivement pour les 1ère, 2ème, 3ème, et 4ème classes de la variable UREE_4CL (ce seront les abscisses des points de la Figure 66 ; les ordonnées étant

les valeurs de la colonne « coefficient » de la Figure 63, les $IC_{95\%}$ étant calculés à partir des SE présentes dans la colonne « Standard Error » de cette même figure).

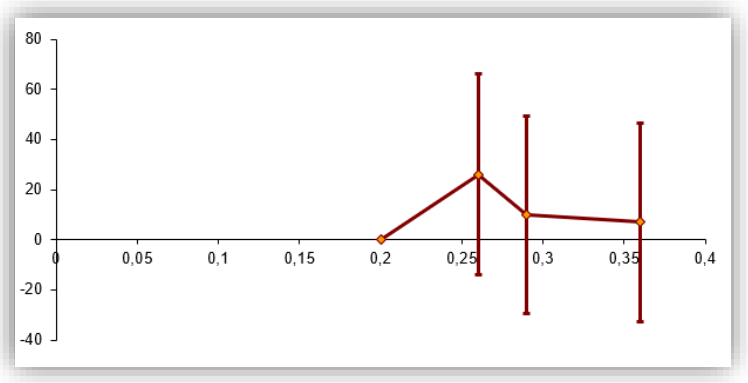


Figure 66

Si l'on n'a aucune raison de penser que l'association entre la concentration en urée et la concentration en ALAT n'est pas linéaire, ou bien si l'on pense qu'il n'y a pas du tout d'association ces deux concentrations, alors la Figure 66 laisserait accepter l'hypothèse de la linéarité de l'association, et le modèle $\overline{ALAT}_{/UREE} = \alpha + \beta.UREE$ fournirait une valeur de β interprétable. Dans le cas contraire, alors vous ne devez pas faire tourner le modèle $\overline{ALAT}_{/UREE} = \alpha + \beta.UREE$, et garder pour vos analyses le modèle ci-dessous, celui qui fournit les résultats de la Figure 63 (et le graphique de la Figure 66).

$\overline{ALAT}_{/(UREE\_4CL)} = \alpha + \beta_{1,2,3}.(UREE\_4CL)$

V. La régression logistique

A. Introduction

Je vous rappelle qu'une régression logistique s'utilise lorsque le CdJ est binaire et non assorti d'un temps de survenue (par exemple, dans une étude cas-témoins ou transversale). Toutes les interprétations des coefficients issus du modèle de régression logistique vont s'appuyer sur celles des coefficients issus du modèle de régression linéaire, avec comme seule différence la suivante : tandis que β représentait la différence de moyennes sur le CdJ quantitatif entre des individus différant de +1 unité sur la variable E dans le modèle de régression linéaire, dans une régression logistique, β représente $Ln(OR_E)$ où l'OR_E quantifie l'association entre E et la présence du CdJ binaire en comparant des individus différant de +1 unité sur E. Par conséquent, vous ne pouvez pas lire la suite de ce guide d'utilisation d'Epi Info si vous n'avez pas lu *tout* ce qui précède dans ce guide (et notamment *tout* ce qui précède sur la régression linéaire) !

Pour l'ensemble des exemples ci-dessous, je vais utiliser comme variable relative au CdJ la variable DECES_3ANS, qui est une variable binaire, valant « 0 » si le chien était toujours en vie 3 ans après l'inclusion dans l'étude, et « 1 » s'il était décédé dans les 3 ans après l'inclusion (je rappelle que tous les chiens avaient été suivis au moins 3 ans, sauf s'ils décédaient avant, et il n'y avait aucun perdu de vue dans l'étude).

B. Interprétation des résultats d'une régression logistique univariée

1. *Modèle de régression logistique avec variable binaire*

Supposons que l'on veuille savoir s'il existe une association brute entre la présence d'un décès dans les 3 ans et le sexe des chiens (variable FEMELLE). Pour répondre à la question, faisons tourner un modèle de régression logistique, qui s'écrit de la façon suivante :

$Logit(\bar{P}_{/FEMELLE}) = \alpha + \beta.FEMELLE$, avec $\bar{P}_{/FEMELLE}$ l'espérance de la probabilité d'être décédé dans les 3 ans selon que le chien est un mâle ou une femelle.

Pour faire tourner ce modèle dans Epi Info, on clique sur « Logistic Regression » (Figure 67.a), on sélectionne la variable Y du modèle, ici DECES_3ANS (Figure 67.b), on sélectionne la variable que l'on veut inclure dans le modèle, ici FEMELLE (Figure 67.c), puis on clique sur « Ok » (Figure 67.d).

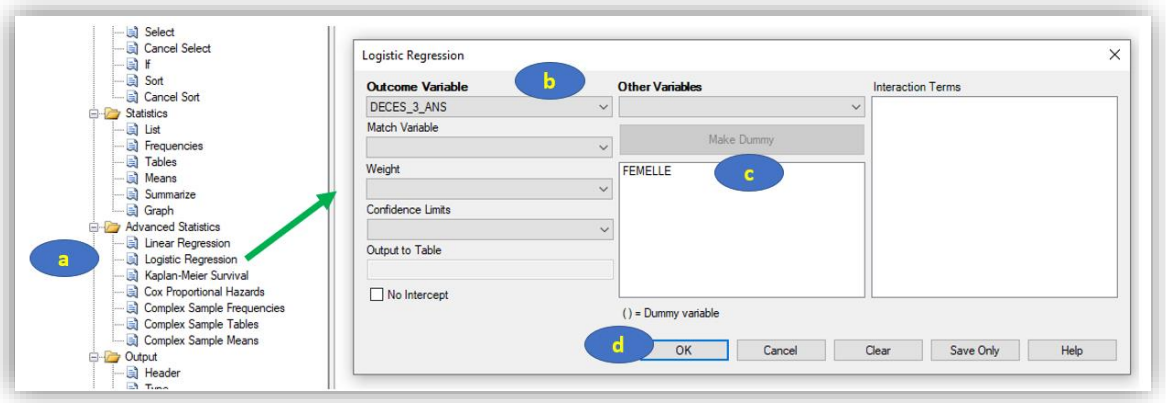


Figure 67

On obtient les résultats présentés sur la Figure 68.

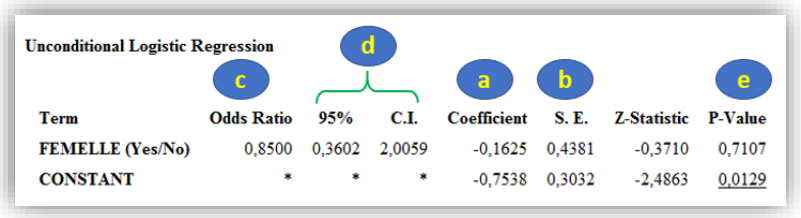


Figure 68

Tout d'abord, la colonne (a) « Coefficient » de la Figure 68 fait référence aux coefficients du modèle de régression logistique, comme dans la régression linéaire. Ainsi, le modèle s'écrit :

$Logit(\bar{P}_{/FEMELLE}) = -0{,}7538 - 0{,}1625.FEMELLE.$

La colonne (b) « S.E. » fait référence à la Standard Error (SE) des coefficients. Comme indiqué dans la Partie II.B de ce Chapitre (page 36), le coefficient β quantifie l'association entre la présence du CdJ (binaire) et E en tant que valeur du $Ln(OR_E)$, où OR_E est l'Odds Ratio quantifiant l'association entre le CdJ et la variable E, pour une augmentation de +1 unité de E. Ici, la variable FEMELLE est binaire, et elle est codée en 0/1 (cf. Chapitre 1, Partie III, page 9), avec « 1 » pour les chiens femelles et « 0 » pour les chiens mâles. Puisque $Ln(OR_{FEMELLE}) = -0{,}1625$, alors l'OR $_{\text{Femelles } versus \text{ mâles}} = e^{-0{,}1625} = 0{,}85$, valeur que l'on retrouve en (c) toujours sur la Figure 68. Dans la mesure où cet OR est < 1, on peut donc dire que, dans l'échantillon, le décès dans les 3 ans était survenu *moins* fréquemment parmi les chiens femelles que parmi les chiens mâles. L'IC$_{95\%}$ de cet OR est [0,36 ; 2,01] (Figure 68.d). Cet IC$_{95\%}$ comprend la valeur « 1 », donc il n'est pas significativement différent de « 1 », ce que l'on retrouve avec un degré

de signification de valeur 0,71 (Figure 68.e), supérieur à 0,05. Ainsi, dans l'échantillon, il n'existait pas d'association significative entre le fait de décéder dans les 3 ans et le sexe des chiens.

2. *Modèle de régression logistique univarié avec variable quantitative*

Supposons que l'on veuille savoir s'il existe une association brute entre la présence d'un décès dans les 3 ans et l'âge des chiens (variable AGE, exprimée en années), à l'aide de la régression logistique.

Je vous rappelle (vivement) que faire tourner le modèle ci-dessous nécessite d'avoir vérifié que l'association entre l'âge et la présence d'un décès à 3 ans est linéaire.

$$Logit\left(\bar{P}_{/AGE}\right) = \alpha + \beta.AGE$$

C'est cette hypothèse que nous allons vérifier désormais, à l'aide de la variable AGE_4CL qui figure déjà dans le fichier de données (cf. Chapitre 1, Partie III, page 9), et dont les classes correspondent aux quartiles (comme je vous recommande de le faire pour vérifier la linéarité d'une association). Nous allons donc faire tourner le modèle ci-dessous :

$$Logit\left(\bar{P}_{/(AGE\_4CL)}\right) = \alpha + \beta_{1,2,3}.(AGE\_4CL)$$

Dans Epi Info, ce modèle ci-dessus correspond à celui-ci-dessous :

$$Logit\left(\bar{P}_{/(AGE\_4CL)}\right) = \alpha + \beta_1.AGE\_4CL(1/0) + \beta_2.AGE\_4CL(2/0) + \beta_3.AGE\_4CL(3/0)$$

Pour cela, après avoir sélectionné la variable AGE_4CL dans Epi Info, il faut cliquer sur « Make Dummy » (cf. Figure 69).

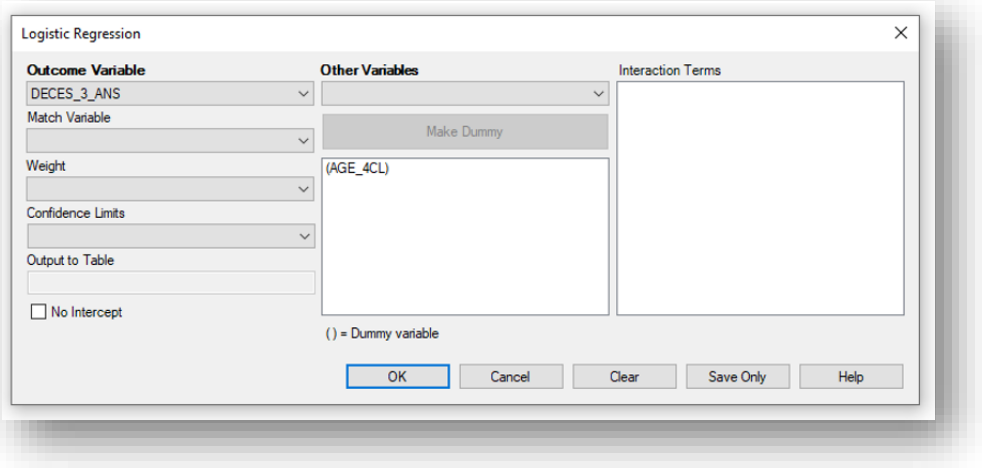


Figure 69

La Figure 70 présente les résultats que l'on obtient après avoir cliqué sur « Ok ».

Unconditional Logistic Regression

Term	Odds Ratio	95%	C.I.	a Coefficient	b S. E.	Z-Statistic	P-Value
AGE_4CL (1/0)	5,0990	0,5575	46,6389	1,6291	1,1293	1,4425	0,1492
AGE_4CL (2/0)	9,9981	1,1500	86,9212	2,3024	1,1034	2,0867	0,0369
AGE_4CL (3/0)	14,7306	1,7175	126,3374	2,6899	1,0965	2,4533	0,0142
CONSTANT	*	*	*	-2,8330	1,0289	-2,7534	0,0059

Figure 70

Pour vérifier la linéarité de l'association entre la présence d'un décès à 3 ans et la variable AGE_4CL, il faut dresser le graphique de la Figure 57 en plaçant quatre points (car AGE_4CL a quatre classes) : le point d'ordonnée 0, le point d'ordonnée β_1, le point d'ordonnée β_2, et le point d'ordonnée β_3. Comme indiqué dans la Partie IV.B.1 de ce Chapitre (page 51), je vous recommande de placer ces quatre points avec comme abscisse la médiane de la variable quantitative (ici, la médiane de AGE) pour chacune des quatre classes de la variable AGE_4CL. Pour cela, nous allons utiliser la commande « MEANS » d'Epi Info (cf. Figure 71).

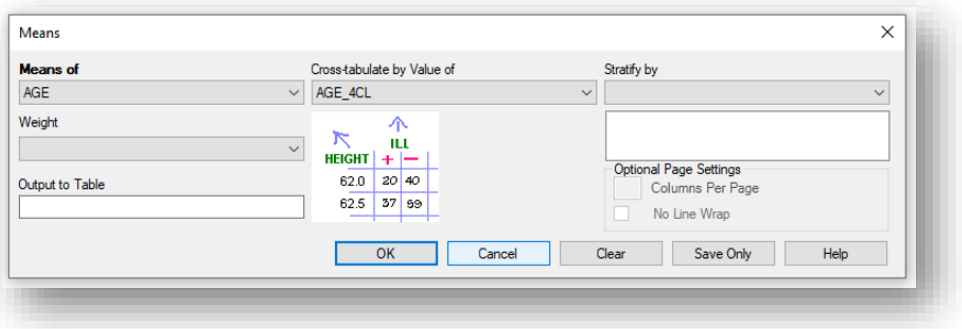
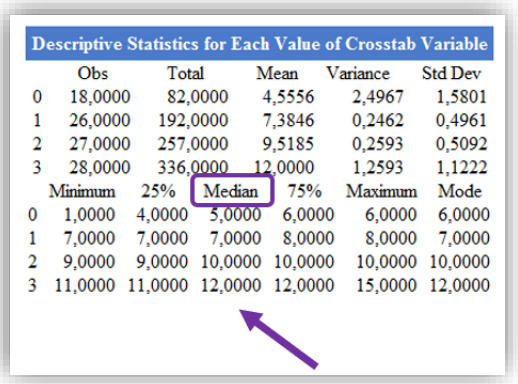


Figure 71

Après avoir cliqué que « Ok », on obtient les résultats ci-dessous (cf. Figure 72).



Descriptive Statistics for Each Value of Crosstab Variable

	Obs	Total	Mean	Variance	Std Dev
0	18,0000	82,0000	4,5556	2,4967	1,5801
1	26,0000	192,0000	7,3846	0,2462	0,4961
2	27,0000	257,0000	9,5185	0,2593	0,5092
3	28,0000	336,0000	12,0000	1,2593	1,1222

	Minimum	25%	Median	75%	Maximum	Mode
0	1,0000	4,0000	5,0000	6,0000	6,0000	6,0000
1	7,0000	7,0000	7,0000	8,0000	8,0000	7,0000
2	9,0000	9,0000	10,0000	10,0000	10,0000	10,0000
3	11,0000	11,0000	12,0000	12,0000	15,0000	12,0000

Figure 72

Le centre des classes de la variable AGE_4CL que nous allons donc utiliser pour dresser le graphique de la Figure 57 sont donc de 5, 7, 10, et 12 respectivement pour les 1ère, 2ème, 3ème, et 4ème classes de la variable AGE_4CL (ce seront les abscisses des points de la Figure 73 ; les ordonnées étant les valeurs de la colonne (a) « coefficient » de la Figure 70, les $IC_{95\%}$ étant calculés à partir des SE présentes dans la colonne (b) « Standard Error » de cette même figure).

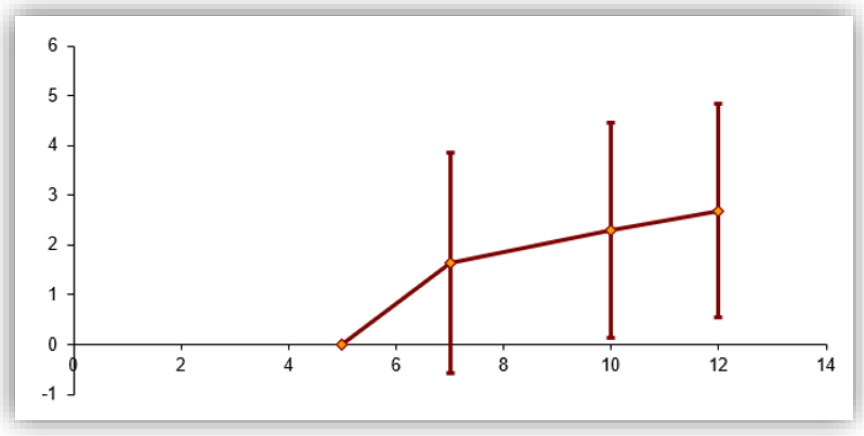


Figure 73

La Figure 73 nous permet d'accepter la linéarité de l'association entre la présence d'un décès à 3 ans et la variable AGE_4CL. Ainsi, on peut aussi accepter la linéarité de l'association entre la présence d'un décès à 3 ans et la variable quantitative AGE. Ainsi, nous pouvons faire tourner le modèle ci-dessous, et interpréter la valeur du coefficient β.

$$Logit(\bar{P}_{/AGE}) = \alpha + \beta.AGE$$

Les résultats fournis par Epi Info après avoir fait tourner ce modèle sont présentés sur la Figure 74.

Unconditional Logistic Regression

Term	Odds Ratio (b)	95% C.I. (c)		Coefficient (a)	S. E.	Z-Statistic	P-Value (d)
AGE	1,3255	1,0988	1,5990	0,2818	0,0957	2,9444	0,0032
CONSTANT	*	*	*	-3,4152	0,9392	-3,6361	0,0003

Figure 74

Le modèle estimé par Epi Info s'écrit donc de la façon suivante (Figure 74.a) :

$$Logit(\bar{P}_{/AGE}) = -3{,}4152 + 0{,}2818.AGE$$

Ainsi, l'OR_{AGE} = $e^{0,2818}$ = 1,33 (Figure 74.b), avec comme $IC_{95\%}$: [1,10 ; 1,60] (Figure 74.c), non significativement différent de « 1 » (Figure 74.d). Nous avons vu précédemment que le coefficient β quantifie l'association entre la présence du CdJ (binaire) et E en tant que valeur du $Ln(OR_E)$, où OR_E est l'Odds Ratio quantifiant l'association entre le CdJ et la variable E, pour une augmentation de +1 unité de E. Ainsi, l'OR_{AGE} de valeur 1,33 s'interprète de la façon suivante : une augmentation de +1 année d'âge pour les chiens de l'échantillon se traduisait par un OR [$IC_{95\%}$] de présenter un décès dans les 3 ans de 1,33 [1,10 ; 1,60]. Cet OR étant significativement différent de « 1 », il existait une association significative entre l'âge des chiens et le fait de présenter un décès dans les 3 ans. Ainsi, voici ce que nous écririons dans un article : « il existait une association significative entre l'âge et le fait de présenter un décès dans les 3 ans (OR [$IC_{95\%}$] pour une augmentation de +1 année d'âge de 1,33 [1,10 ; 1,60], $p < 0{,}01$) ».

3. *Modèle de régression logistique univarié avec variable qualitative ordinale*

Supposons que l'on veuille savoir s'il existe une association brute entre la présence d'un décès dans les 3 ans et la cholestérolémie des chiens (hypocholestérolémie, normocholestérolémie, et hypercholestérolémie, en utilisant la variable CHOLES_3CL), à l'aide de la régression logistique.

Je vous rappelle (encore et toujours) que faire tourner le modèle ci-dessous nécessite d'avoir vérifié que l'association entre la cholestérolémie (en trois classes) et la présence d'un décès à 3 ans est linéaire.

$$Logit\left(\bar{P}_{/CHOLES\_3CL}\right) = \alpha + \beta.CHOLES\_3CL$$

C'est ce nous allons vérifier désormais. Nous allons donc faire tourner le modèle ci-dessous :

$$Logit\left(\bar{P}_{/(CHOLES\_3CL)}\right) = \alpha + \beta_{1,2}.(CHOLES\_3CL)$$

Dans Epi Info, le modèle ci-dessus correspond à celui-ci-dessous :

$$Logit\left(\bar{P}_{/(CHOLES\_3CL)}\right) = \alpha + \beta_1.CHOLES\_3CL(1/0) + \beta_2.CHOLES\_3CL(2/0)$$

Pour faire tourner ce modèle, après avoir sélectionné la variable CHOLES_3CL dans Epi Info, il faut cliquer sur « Make Dummy » (cf. Figure 75).

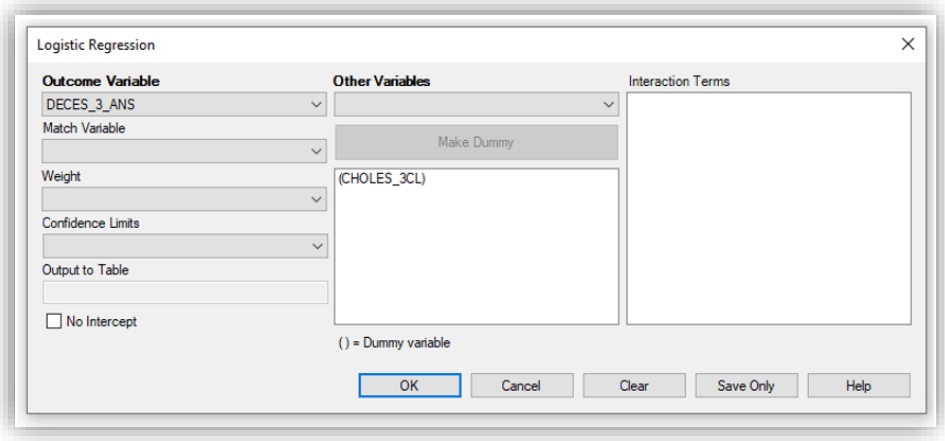


Figure 75

La Figure 76 présente les résultats que l'on obtient après avoir cliqué sur « Ok ».

Unconditional Logistic Regression

Term	Odds Ratio	95%	C.I.	Coefficient (a)	S. E. (b)	Z-Statistic	P-Value
CHOLES_3CL (1/0)	0,6877	0,2373	1,9935	-0,3744	0,5430	-0,6895	0,4905
CHOLES_3CL (2/0)	2,1772	0,6807	6,9634	0,7780	0,5932	1,3116	0,1896
CONSTANT	*	*	*	-0,8650	0,4215	-2,0524	0,0401

Figure 76

Pour vérifier la linéarité de l'association entre la présence d'un décès à 3 ans et la variable CHOLES_3CL, il faut dresser le graphique de la Figure 57 en plaçant trois points (car CHOLES_3CL a trois classes) : le point d'ordonnée 0, le point d'ordonnée β_1, et le point d'ordonnée β_2 (cf. valeurs dans la colonne (a) de la Figure 76) et comme abscisse les valeurs de « 0 », « 1 », et « 2 » qui sont les valeurs des trois classes de CHOLES_3CL. A partir des valeurs des SE des coefficients β (colonne (b) de la Figure 76) qui permettent de calculer les $IC_{95\%}$ des coefficients, on dresse le graphique présenté sur la Figure 77.

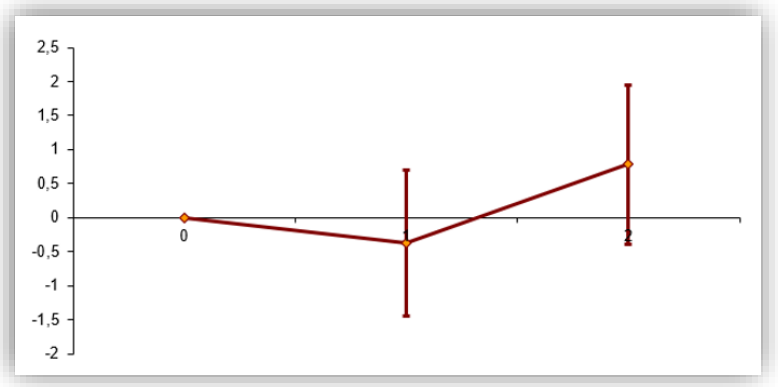


Figure 77

A partir d'un tel graphique, il pourrait sembler difficile d'accepter la linéarité de l'association entre la présence d'un décès à 3 ans et la cholestérolémie (en trois classes). Attention, si vous décidez de regrouper deux (ou plusieurs) classes en fonction des résultats, vous n'aurez pas le droit d'inférer avec autant de conviction que si vous aviez déjà décidé de regrouper certaines classes *avant* d'avoir vu les données. A partir du graphique ci-dessus, nous pourrions être tentés de regrouper les deux classes « hypocholestérolémie » et « normocholestérolémie ». En allant au bout, nous créerions alors une variable binaire à partir de la variable CHOLES_3CL, qui vaudrait « 0 » pour les chiens avec une hypocholestérolémie ou une normocholestérolémie, et « 1 » dans le cas d'hypercholestérolémie. Il se trouve que cette variable a déjà été créée dans le fichier de données (cf. Chapitre 1, Partie III, page 9), qui s'appelle « HYPER_CHOLES ». Le modèle correspondant est celui-ci-dessous :

$$Logit\left(\bar{P}_{/HYPER\_CHOLES}\right) = \alpha + \beta.HYPER\_CHOLES$$

Les résultats du modèle, que l'on fait tourner dans Epi Info, sont présentés sur la Figure 78 ci-dessous.

Unconditional Logistic Regression

Term	Odds Ratio	95%	C.I.	Coefficient	S. E.	Z-Statistic	P-Value
HYPER_CHOLES (Yes/No)	2,7500	1,0435	7,2469	1,0116	0,4944	2,0462	0,0407
CONSTANT	*	*	*	-1,0986	0,2649	-4,1470	0,0000

Figure 78

Ainsi, dans la mesure où l'OR, de valeur 2,75, est supérieur à « 1 », on peut dire que dans l'échantillon, le décès dans les 3 ans était plus fréquent parmi les chiens qui avaient présenté une hypercholestérolémie que parmi les autres chiens (OR $_{\text{hypercholestérolémie } \textit{versus} \text{ pas d'hypercholestérolémie}}$ = 2,75 [1,04 ; 7,25]$_{95\%}$; p = 0,04). Il existait donc une association significative entre la présence d'un décès dans les 3 ans et la présence d'une hypercholestérolémie. Cela dit, comme je l'ai déjà écrit, cette association significative est le résultat d'un regroupement *au vu des données*. Par conséquent, il n'est pas possible d'inférer (ou alors, faites-le vos risques et périls) que dans la population des chiens adultes, il y a des chances pour qu'il existe une association réelle entre la présence d'un décès dans les 3 ans et celle d'une hypercholestérolémie.

4. *Modèle de régression logistique univarié avec variable qualitative nominale*

Supposons que l'on veuille savoir s'il existe une association brute entre la présence d'un décès dans les 3 ans et la race des chiens (en utilisant la variable RACE_4CL), à l'aide de la régression logistique.

Nous allons reprendre exactement la même démarche que celle avec la régression linéaire, à l'époque où nous voulions savoir s'il existait une association entre la concentration en ALAT et la race des chiens (cf. Chapitre 3, Partie III.B.5.c, page 46). Le modèle que nous allons donc faire tourner est celui-ci-dessous :

$$Logit\left(\bar{P}_{/(RACE\_4CL)}\right) = \alpha + \beta_{1,2,3}.(RACE\_4CL)$$

Dans Epi Info, le modèle ci-dessus correspond à celui-ci-dessous :

$$Logit\left(\bar{P}_{/(RACE\_4CL)}\right) = \alpha + \beta_1.RACE\_4CL(1/0) + \beta_2.RACE\_4CL(2/0) + \beta_3.RACE\_4CL(3/0)$$

Ainsi (cf. Figure 79), on sélectionne la variable Y du modèle, ici DECES_3ANS (a), on sélectionne la variable que l'on veut inclure dans le modèle sous forme de « dummy variables », ici RACE_4CL (b), on clique sur « RACE_4CL », ce qui fait apparaître « Make Dummy » (c). Après avoir cliqué sur « Make Dummy », on voit des parenthèses apparaître autour de « RACE_4CL » (d), ce qui indique bien que des « dummy variables » vont être créées par Epi Info au moment de faire tourner le modèle. Puis on clique sur « Ok » (e).

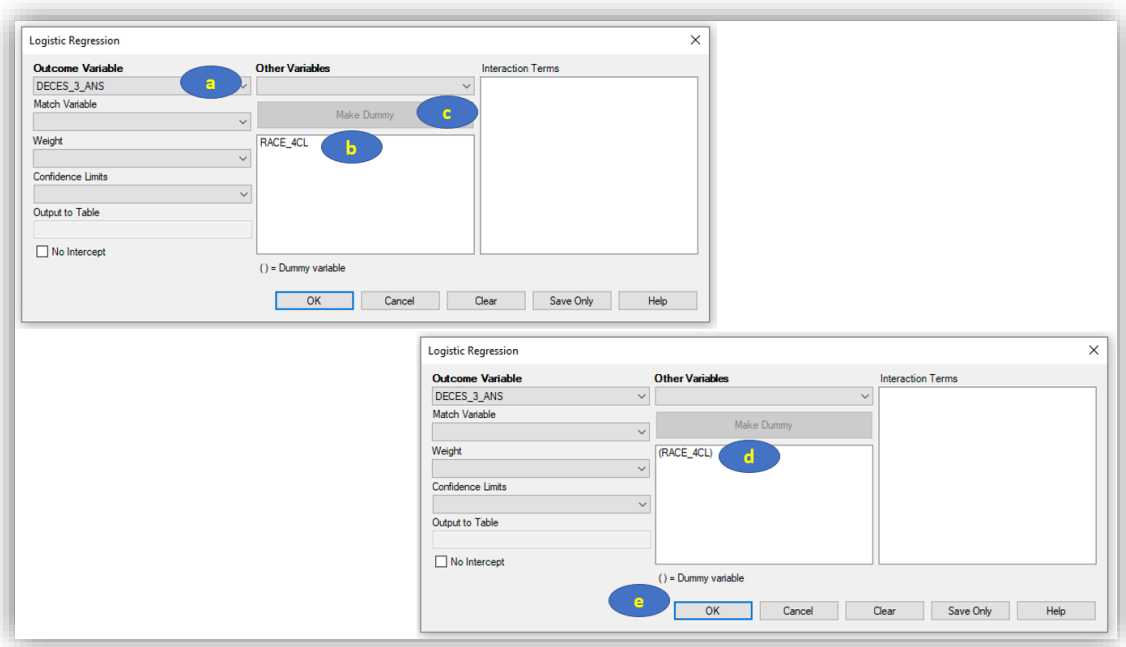
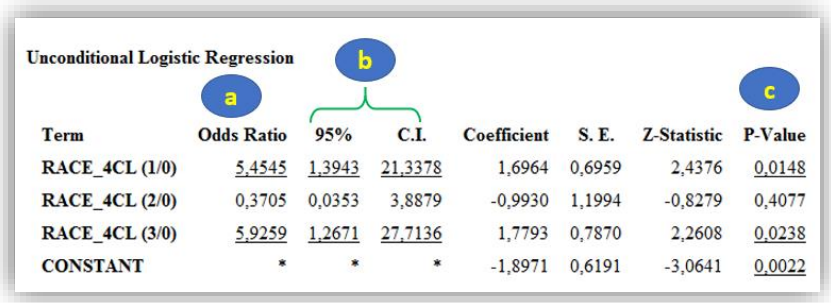


Figure 79

On obtient les résultats présentés sur la Figure 80 ci-dessous.



Unconditional Logistic Regression

Term	Odds Ratio	95% C.I.		Coefficient	S. E.	Z-Statistic	P-Value
RACE_4CL (1/0)	5,4545	1,3943	21,3378	1,6964	0,6959	2,4376	0,0148
RACE_4CL (2/0)	0,3705	0,0353	3,8879	-0,9930	1,1994	-0,8279	0,4077
RACE_4CL (3/0)	5,9259	1,2671	27,7136	1,7793	0,7870	2,2608	0,0238
CONSTANT	*	*	*	-1,8971	0,6191	-3,0641	0,0022

Figure 80

Les résultats présentés sur la Figure 80 s'interprètent de la façon suivante. Le décès à 3 ans était significativement plus fréquent parmi les chiens de race Labrador (classe « 1 » pour RACE_4CL) que parmi les chiens de race Golden, la classe « 0 » pour RACE_4CL, classe dite de référence (OR $_{\text{Labrador } versus \text{ Golden}}$ = 5,45 [1,39 ; 21,34]$_{95\%}$; p = 0,01) ; il était aussi significativement plus fréquent parmi les chiens d'autres race (classe « 3 » pour RACE_4CL) que parmi les chiens de race Golden (OR $_{\text{Autre race } versus \text{ Golden}}$ = 5,29 [1,27 ; 27,71]$_{95\%}$; p = 0,02). Il était en revanche moins fréquent parmi les chiens de race croisée Golden/Labrador (classe « 2 » pour RACE_4CL) que parmi les chiens de race Golden (OR $_{\text{Race croisée } versus \text{ Golden}}$ = 0,37 [0,04 ; 3,89]$_{95\%}$; p = 0,41), sans que cette différence de fréquence ne soit significative.

C. Interprétation des résultats d'une régression logistique multivariée

Supposons que l'on souhaite étudier l'association entre la présence d'un décès à 3 ans et le sexe des chiens, ajustée sur l'âge et la race des chiens. Le modèle de régression logistique multivarié correspondant est le suivant :

$$Logit\left(\bar{P}_{/FEMELLE,AGE,(RACE\_4CL)}\right) = \alpha + \beta.FEMELLE + \gamma.AGE + \delta_{1,2,3}.(RACE\_4CL)$$

Je rappelle (encore et encore) que tous les coefficients du modèle (β, γ, δ_1, δ_2, et δ_3) ne sont interprétables que si le modèle repose sur l'hypothèse que l'association entre l'âge et la présence d'un décès à 3 ans est linéaire. Cela dit, nous avons vu que, de façon brute (c'est-à-dire, non ajustée), l'association entre l'âge et la présence d'un décès à 3 ans pouvait être considérée comme linéaire (cf. Figure 73). Même s'il est toujours possible qu'un biais de confusion puisse modifier la forme de l'association entre une variable quantitative et le CdJ, on peut considérer que si l'association brute peut être considérée comme linéaire entre le CdJ et la variable quantitative, alors elle le restera après ajustement sur d'autres variables dans un modèle de régression multivarié. Donc, les estimations α, β, γ, δ_1, δ_2, et δ_3) du modèle ci-dessus seront interprétables.

Pour faire tourner le modèle ci-dessus, voici ce qu'il faut sélectionner dans Epi Info (cf. Figure 81) :

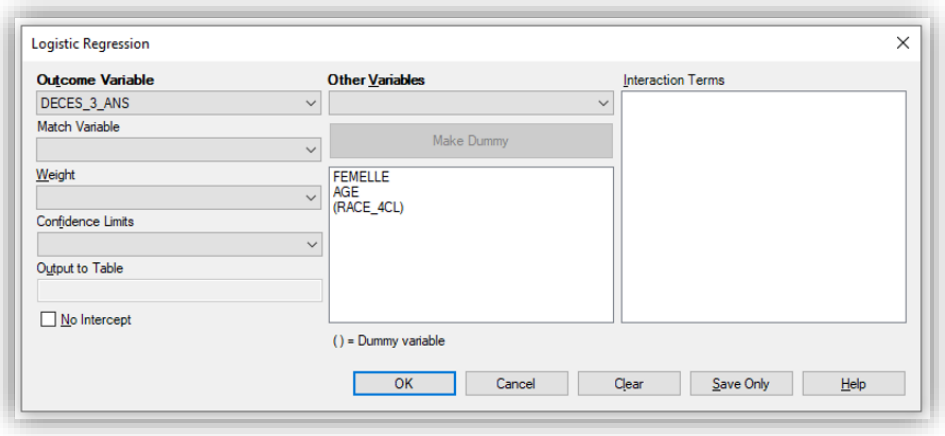


Figure 81

On obtient les résultats présentés sur la Figure 82.

Unconditional Logistic Regression

	b	c		a			d
Term	**Odds Ratio**	**95%**	**C.I.**	**Coefficient**	**S. E.**	**Z-Statistic**	**P-Value**
FEMELLE (Yes/No)	0,8339	0,3007	2,3126	-0,1816	0,5204	-0,3490	0,7271
AGE	1,4357	1,1431	1,8032	0,3617	0,1163	3,1105	0,0019
RACE_4CL (1/0)	4,9093	1,1504	20,9509	1,5911	0,7403	2,1492	0,0316
RACE_4CL (2/0)	0,1753	0,0147	2,0834	-1,7414	1,2630	-1,3788	0,1680
RACE_4CL (3/0)	4,1598	0,8191	21,1246	1,4255	0,8291	1,7193	0,0856
CONSTANT	*	*	*	-4,9057	1,2777	-3,8396	0,0001

Figure 82

A partir des résultats de la régression logistique présentés sur la Figure 82, ce modèle estimé par Epi Info reliant la présence d'un décès à 3 ans au sexe, à l'âge, et à la race, s'écrit de la façon suivante (à partir des valeurs de la colonne (a) de la Figure 82) :

$$Logit\left(\bar{P}_{/FEMELLE,AGE,(RACE\_4CL)}\right) = -4{,}9057 - 0{,}1816.FEMELLE + 0{,}3617.AGE + 1{,}5911.RACE\_4CL(1/0) - 1{,}7414.RACE\_4CL(2/0) + 1{,}4255.RACE\_4CL(3/0)$$

J'insiste à nouveau sur le fait que vous devez impérativement avoir conscience que dès que l'on inclut dans un modèle une variable qualitative ordinale ou quantitative telle quelle (c'est-à-dire, sans l'inclure sous forme de « dummy variables »), *tous* les coefficients du modèle (y compris ceux associés à des variables binaires) ne sont interprétables que si l'hypothèse de la linéarité de l'association pour chacune des variables qualitatives ordinales ou quantitatives incluses dans le modèle est vérifiée. Le modèle ci-dessus repose donc sur l'hypothèse de la linéarité de l'association entre la présence d'un décès à 3 ans et l'âge. Cette hypothèse est vérifiée (cf. raisonnement ci-dessus).

Dans une régression logistique, comme nous l'avons vu précédemment, les valeurs des coefficients ne sont pas directement interprétables. Nous allons donc directement interpréter les valeurs des OR (Figure 82.b), assortis de leur $IC_{95\%}$ (Figure 82.c) et de leur degré de signification (Figure 82.d). Ainsi, indépendamment de l'âge et de la race, dans l'échantillon, il n'existait pas d'association significative entre le sexe des chiens et la présence d'un décès à 3 ans ($OR_{\text{femelles } versus \text{ mâles}}$ = 0,83 $[0{,}30 ; 2{,}31]_{95\%}$; p = 0,73). Indépendamment du sexe des chiens et de la race, dans l'échantillon, il existait une association significative entre l'âge des chiens et la présence d'un décès à 3 ans ($OR_{\text{pour une augmentation de +1 année d'âge}}$ =

1,44 [1,14 ; 1,80]$_{95\%}$; p < 0,01). Indépendamment du sexe et de l'âge des chiens, dans l'échantillon, le décès à 3 ans était significativement plus fréquent parmi les chiens de race Labrador (classe « 1 » pour RACE_4CL) que parmi les chiens de race Golden (classe « 0 » pour RACE_4CL, qui est la classe de référence ; OR $_{\text{Labrador } versus \text{ Golden}}$ = 4,91 [1,15 ; 20,95]$_{95\%}$; p = 0,03). Indépendamment du sexe et de l'âge des chiens, dans l'échantillon, il n'existait pas de différence significative de fréquence d'un décès à 3 ans ni entre les chiens de race croisée Golden/Labrador (classe « 2 » pour RACE_4CL) et les chiens de race Golden (OR $_{\text{Race croisée } versus \text{ Golden}}$ = 0,18 [0,01 ; 2,08]$_{95\%}$; p = 0,18), ni entre les chiens d'autre race (classe « 3 » pour RACE_4CL) et les chiens de race Golden (OR $_{\text{Autre race } versus \text{ Golden}}$ = 4,16 [0,82 ; 21,12]$_{95\%}$; p = 0,09).

VI. Le modèle (à risques proportionnels) de Cox

A. Introduction

Vous devez avoir lu le polycopié d'analyse de survie et (bien entendu) tout ce qui précède dans ce guide avant de poursuivre.

Si le CdJ est binaire et assorti d'un temps de survenue (par exemple, dans une étude de cohorte), alors le modèle de régression est le modèle de Cox (Cox, 1972). Le modèle de Cox s'écrit :

$Ln\left(\overline{\lambda(t)}_{/E_1,E_2,\ldots,E_N}\right) = Ln\left(\lambda_0(t)\right) + \sum_{i=1}^{N} \beta_i . E_i$, où « $\overline{\lambda(t)}_{/E_1,E_2,\ldots,E_N}$ » est l'espérance de l'incidence instantanée du CdJ en fonction de la valeur des variables E_i incluses dans le modèle. De façon générale, β_i quantifie l'association entre la *survenue* du CdJ (binaire) et E_i en tant que valeur du $Ln\left(HR_{E_i}\right)$, où HR_{E_i} est le Risque Relatif (« Hazard Ratio ») quantifiant l'association entre la survenue du CdJ et la variable E_i.

A part le fait que le modèle de Cox fournit un HR alors que la régression logistique fournit un OR (et à part le fait que le modèle de Cox repose sur une hypothèse qui s'appelle l'hypothèse de la proportionnalité des risques, cf. Partie VI.D dans ce chapitre), tout ce que j'ai écrit pour la régression logistique est valable pour le modèle de Cox. Ainsi, je vais passer beaucoup moins de temps sur le modèle de Cox que je n'en ai passé sur la régression logistique.

B. Interprétation des résultats d'un modèle de Cox univarié

Comme je l'ai écrit ci-dessous, la démarche d'utilisation d'un modèle de Cox est identique à celle d'une régression logistique. Par conséquent, dans cette partie « modèle de Cox univarié », je ne vais pas détailler selon les types de variables incluses dans le modèle (binaire, quantitative, qualitative ordinale, et qualitative nominale). Je vais uniquement présenter la démarche pour une variable binaire, et notamment pour la variable DEMARCHE_ANORMALE.

Supposons que l'on veuille savoir s'il existe une association brute entre la survenue d'un décès au cours du suivi et la démarche des chiens (anormale *versus* normale, en utilisant la variable DEMARCHE_ANORMALE) en quantifiant cette association. Nous avions déjà effectué cette analyse à l'aide des courbes de Kaplan-Meier (Figure 38), mais les courbes de Kaplan-Meier ne permettent pas de quantifier une association. Elles permettent « simplement » de représenter visuellement une association (ce qui est souvent nécessaire) et de tester statistiquement (à l'aide du test du log-rank) l'association.

Pour répondre à la question, faisons tourner le modèle de Cox univarié, qui s'écrit de la façon suivante :

$Ln\left(\overline{\lambda(t)}_{/DEMARCHE\_ANORMALE}\right) = Ln\left(\lambda_0(t)\right) + \beta . DEMARCHE\_ANORMALE$, où « $\overline{\lambda(t)}_{/DEMARCHE\_ANORMALE}$ » est l'espérance de l'incidence instantanée du décès selon que les chiens ont ou n'ont pas de démarche anormale.

Pour faire tourner ce modèle dans Epi Info, on clique sur « Cox Proportional Hazards » (Figure 83.a), on sélectionne la variable relative à l'événement (ici DECES ; Figure 83.b), la variable que prend cette variable pour les individus qui présentent l'événement (ici « 1 » ; Figure 83.c), et la variable relative au temps de survie (ici SURVIE ; Figure 83.d). Attention, si le modèle de Cox est univarié, il faut sélectionner la variable relative à l'exposition sous « Test Group Variable » (Figure 83.e) et non pas sous « Predictor Variables » (Figure 83.f). Puis on clique sur « Ok » (Figure 83.g).

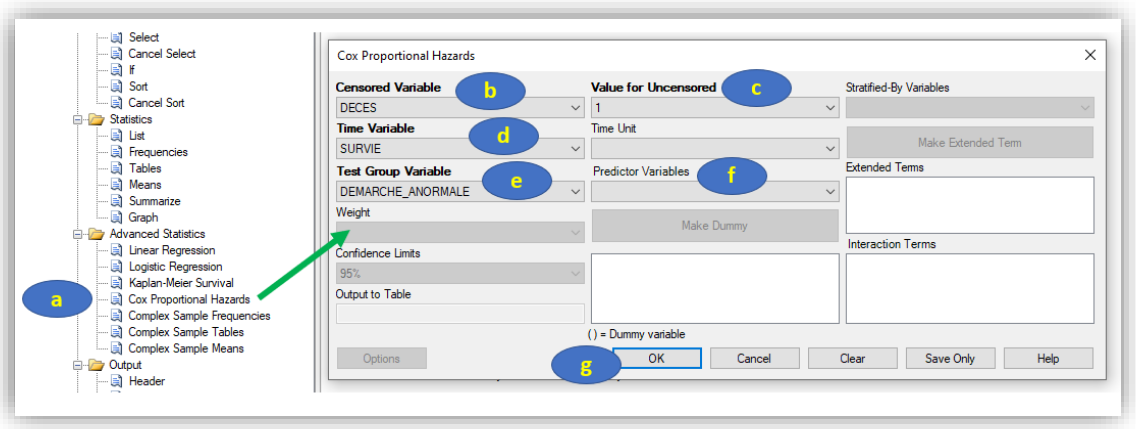


Figure 83

La Figure 84 présente les résultats que l'on obtient après avoir cliqué sur « Ok ».

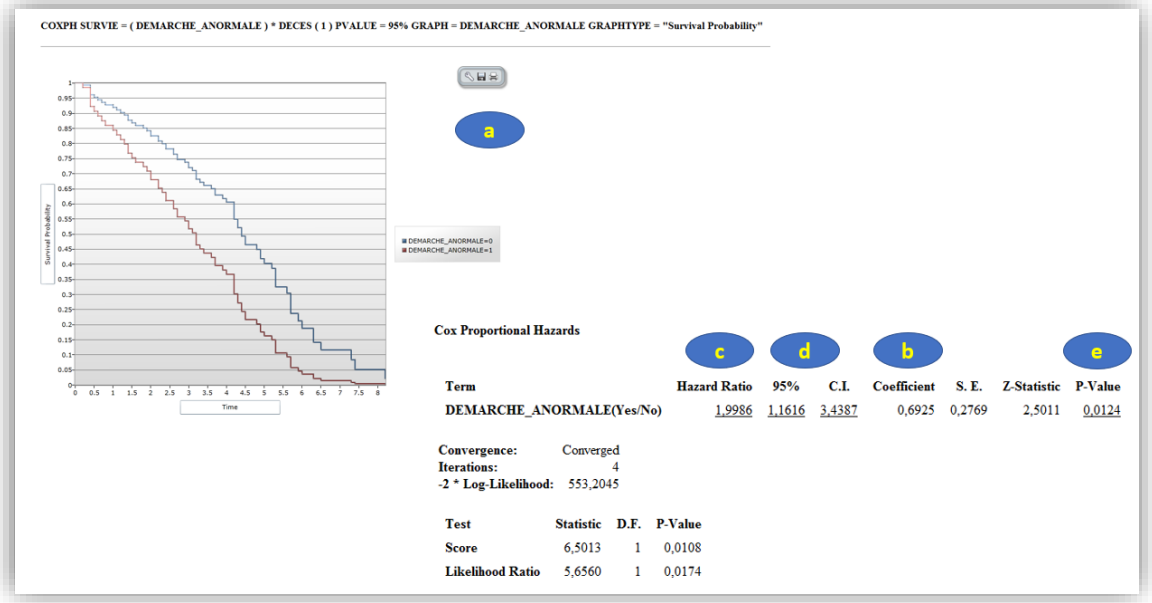


Figure 84

Tout d'abord, Epi Info fournit la courbe de Kaplan-Meier représentant l'association entre la présence d'une démarche anormale et la survenue d'un décès au cours du suivi (sans que le degré de signification du test du log-rank ne soit cependant fourni ; Figure 84.a). Epi Info fournit la valeur de β du modèle ci-dessus, de valeur 0,6925 (Figure 84.b). De façon générale, dans un modèle de Cox, le coefficient β quantifie l'association entre la survenue du CdJ et E en tant que valeur du $Ln(HR_E)$, où HR_E est l'HR quantifiant l'association entre le CdJ et la variable E, pour une augmentation de +1 unité de E. Ici, la variable DEMARCHE_ANORMALE est binaire, et elle est codée en 0/1 (cf. Chapitre 1, Partie III, page 9), avec « 1 » pour les chiens avec une démarche anormale et « 0 » pour les chiens avec une démarche normale. Puisque $Ln(HR_{DEMARCHE\_ANORMALE}) = 0{,}6925$, alors l'HR Démarche anormale *versus* normale

= $e^{0,6925}$ = 2,00, valeur que l'on retrouve en (c) toujours sur la Figure 84. L'$IC_{95\%}$ de ce $HR_{\text{Démarche anormale } \textit{versus} \text{ normale}}$ est [1,16 ; 3,44] (Figure 84.d), et ce HR est significativement différent de 1 (valeur du degré de signification égale à 0,01 ; Figure 84.e). Dans l'échantillon, il n'existait donc pas de différence significative de survenue d'un décès entre les chiens présentant une démarche anormale et les chiens présentant une démarche normale.

C. Interprétation des résultats d'un modèle de Cox multivarié

Supposons que l'on souhaite étudier l'association entre la survenue d'un décès et la démarche des chiens, ajustée sur l'âge et la race des chiens. Le modèle de Cox multivarié correspondant est donc le suivant :

$$Ln\left(\overline{\lambda(t)}_{/DEMARCHE\_ANORMALE,AGE,(RACE\_4CL)}\right) = Ln\left(\lambda_0(t)\right) + \beta.DEMARCHE\_ANORMALE + \gamma.AGE + \delta_{1,2,3}.(RACE\_4CL)$$

Je rappelle (à nouveau) que tous les coefficients du modèle (β, γ, δ_1, δ_2, et δ_3) ne sont interprétables que si le modèle repose sur l'hypothèse que l'association entre l'âge et la survenue d'un décès est linéaire (ce dont j'ai déjà parlé pour la régression linéaire et la régression logistique). Ainsi, avant de faire tourner un tel modèle, il aurait fallu vérifier cette linéarité de l'association avec l'âge dans un modèle de Cox brut incluant les « dummy variables » relatives à la variable AGE_4CL. (La démarche avec le modèle de Cox est identique à celle déjà présentée pour la régression linéaire et la régression logistique ; pour la régression logistique, cf. Chapitre 3, Parties V.B.2, page 60, et V.B.3, page 63). Nous allons faire l'hypothèse que l'association avec l'âge est linéaire. Donc, les estimations α, β, γ, δ_1, δ_2, et δ_3) du modèle ci-dessus seront interprétables.

Pour faire tourner le modèle ci-dessus, je vous recommande de ne pas cliquer sur « Cox Proportional Hazards » comme nous l'avions fait précédemment (cf. Figure 83.a) car Epi Info plante dans certaines situations. Je vous recommande de taper votre programme dans l'éditeur de programme. Voici la syntaxe générale (ce qui est en italique doit être remplacer par vous) :

```
COXPH TEMPS_SURVIE = VAR1 VAR2 … VARk * VAR_EVENT (VALEUR_EVENT)
```

Avec « *TEMPS_SURVIE* » la variable relative au temps de survie, « *VAR1 VAR2 … VARk* » la liste des *k* variables, séparées par un espace, que l'on souhaite inclure dans le modèle de Cox multivarié, « *VAR_EVENT* » la variable relative à l'événement dans l'analyse de survie, et « *VALEUR_EVENT* » la valeur numérique correspondant aux individus qui ont présenté l'événement au cours du suivi. Si une des variables est qualitative nominale, ou si une variable est qualitative ordinale mais ne vérifiant la linéarité de l'association, il faudra taper le nom de la variable en le mettant entre parenthèses. Ainsi, pour faire tourner le modèle ci-dessus, il faudra taper dans l'éditeur de programme (cf. Figure 85) :

```
COXPH SURVIE = DEMARCHE_ANORMALE AGE (RACE_4CL) * DECES (1)
```

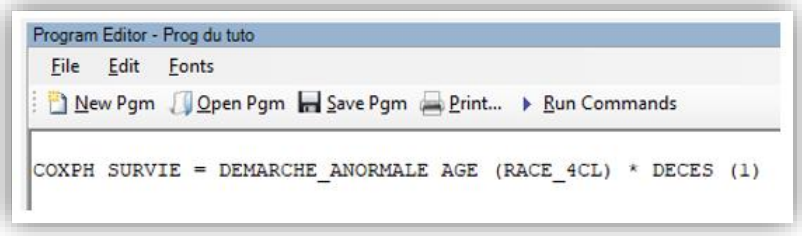


Figure 85

Ensuite, on sélectionne la ligne entière avec la souris, puis on clique sur « Run Commands » (cf. Figure 86.a).

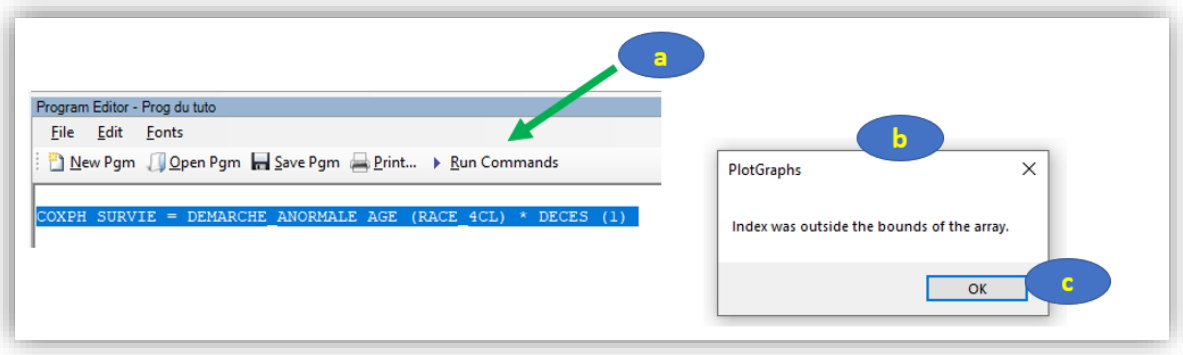


Figure 86

Une fenêtre pop-up apparaît (Figure 86.b), ce n'est pas grave, puis on clique sur « Ok » (Figure 86.c). Voici les résultats que l'on obtient (Figure 87) :

Cox Proportional Hazards

b c a d

Term	Hazard Ratio	95%	C.I.	Coefficient	S. E.	Z-Statistic	P-Value
DEMARCHE_ANORMALE	2,2223	1,2755	3,8719	0,7985	0,2833	2,8190	0,0048
AGE	1,2673	1,1392	1,4097	0,2369	0,0544	4,3582	0,0000
RACE_4CL(1/0)	1,7597	0,8925	3,4694	0,5651	0,3464	1,6317	0,1028
RACE_4CL(2/0)	1,1262	0,5143	2,4662	0,1188	0,3999	0,2971	0,7664
RACE_4CL(3/0)	1,3504	0,6036	3,0208	0,3004	0,4108	0,7312	0,4647

Figure 87

A partir des résultats du modèle de Cox présentés sur la Figure 87, ce modèle estimé par Epi Info reliant la survenue d'un décès à la démarche anormale, à l'âge, et à la race, s'écrit de la façon suivante (valeurs situées dans la colonne (a) de la Figure 87) :

$$Ln\left(\overline{\lambda(t)}_{/DEMARCHE\_ANORMALE,AGE,(RACE\_4CL)}\right) = Ln\left(\lambda_0(t)\right) + 0{,}7985.DEMARCHE\_ANORMALE + 0{,}2369.AGE + 0{,}5651.RACE\_4CL(1/0) + 0{,}1188.RACE\_4CL(2/0) + 0{,}3004.RACE\_4CL(3/0)$$

J'insiste à nouveau sur le fait que vous devez impérativement avoir conscience que dès que l'on inclut dans un modèle une variable qualitative ordinale ou quantitative telle quelle (c'est-à-dire, sans l'inclure sous forme de « dummy variables »), *tous* les coefficients du modèle (y compris ceux associés à des variables binaires) ne sont interprétables que si l'hypothèse de la linéarité de l'association pour chacune des variables qualitatives ordinales ou quantitatives incluses dans le modèle est vérifiée. Le modèle ci-dessus repose donc sur l'hypothèse de la linéarité de l'association entre la survenue d'un décès et l'âge. Nous partirons du principe que cette hypothèse est vérifiée (la démarche de vérification est exactement celle décrite page 51).

Dans un modèle de Cox, comme nous l'avons vu précédemment dans la régression logistique, les valeurs des coefficients ne sont pas directement interprétables. Nous allons donc directement interpréter les valeurs des HR (Figure 87.b), assortis de leur $IC_{95\%}$ (Figure 87.c) et de leur degré de signification (Figure 87.d).

Ainsi, indépendamment de l'âge et de la race, dans l'échantillon, il existait une association significative entre la démarche des chiens (anormale *versus* normale) et la survenue d'un décès ($HR_{\text{Démarche anormale } versus \text{ normale}}$ = 2,22 $[1{,}28 ; 3{,}87]_{95\%}$; $p < 0{,}01$). Indépendamment de la démarche des chiens et de leur race, dans l'échantillon, il existait une association significative entre l'âge des chiens et la survenue

d'un décès ($HR_{\text{pour une augmentation de +1 année d'âge}} = 1,27\ [1,14 ; 1,41]_{95\%}$; $p < 0,01$). Indépendamment de la démarche des chiens et de leur âge, dans l'échantillon, il n'existait de différence significative de survenue d'un décès ni entre les chiens de race Labrador (classe « 1 » pour RACE_4CL) et les chiens de race Golden (classe « 0 » pour RACE_4CL ; $HR_{\text{Labrador } versus \text{ Golden}} = 1,76\ [0,89 ; 3,47]_{95\%}$; $p = 0,10$), ni entre les chiens de race croisée Golden/Labrador (classe « 2 » pour RACE_4CL) et les chiens de race Golden (classe « 0 » pour RACE_4CL ; $HR_{\text{Race croisée } versus \text{ Golden}} = 1,13\ [0,51 ; 1,47]_{95\%}$; $p = 0,77$), ni entre les chiens d'autre race (classe « 3 » pour RACE_4CL) et les chiens de race Golden (classe « 0 » pour RACE_4CL ; $HR_{\text{Autre race } versus \text{ Golden}} = 1,35\ [0,60 ; 3,02]_{95\%}$; $p = 0,47$).

D. Vérification de l'hypothèse de la proportionnalité des risques

1. Introduction

Le modèle de Cox repose sur l'hypothèse de la proportionnalité des risques (HRP). Plus spécifiquement, chaque variable incluse dans un modèle de Cox doit vérifier l'HRP. Cette hypothèse pour une exposition E est vérifiée si et seulement si la valeur du HR_E est constante au cours du temps, au sein de la population cible. Supposons une exposition binaire E, avec des animaux exposés à E (E+) et non exposés à E (E-). La valeur du HR_E ne serait pas constante au cours du temps s'il se passait, par exemple, la chose suivante, dans la population cible : peu de temps après J0, les animaux E+ sont deux fois plus à risque de survenue du CdJ que les animaux E- mais plus longtemps après J0, les animaux E+ ne sont finalement pas plus à risque de survenue du CdJ que les animaux E-. Cette situation peut arriver lorsque l'exposition étudiée a un effet seulement à court terme sur la survenue du CdJ (peu de temps après J0). La situation où l'effet de l'exposition E étudiée n'existe qu'à long terme (pas d'effet peu de temps après J0, mais un effet qui commence à exister longtemps après J0) est aussi une situation où l'HRP ne serait pas vérifiée pour E. Cette hypothèse HRP doit être vérifiée pour chaque variable incluse dans un modèle de Cox.

2. Vérification avec Epi Info

Il existe de nombreuses façons de vérifier l'HRP (Hess, 1995). Epi Info ne permet de ne la vérifier que d'une seule façon, et ce n'est pas la plus aisée pour la vérifier !... Il s'agit de vérifier l'HRP au vu des courbes de Kaplan-Meier. Dans tous les cas, cette vérification utilise les données de l'échantillon. Il faut toujours garder en tête qu'une HRP peut ne pas être vérifiée dans l'échantillon alors qu'elle l'est dans la population cible à cause de la fluctuation d'échantillonnage (l'inverse est tout aussi vrai). Sauf que dans la mesure où nous n'aurons jamais accès aux données de la population cible (seulement par notre intuition « médicale »), nous ne disposons que des données de l'échantillon pour vérifier cette hypothèse.

Les deux courbes de Kaplan-Meier ci-dessous révèlent clairement une HRP non vérifiée pour la variable relative au groupe (cf. Figure 88).

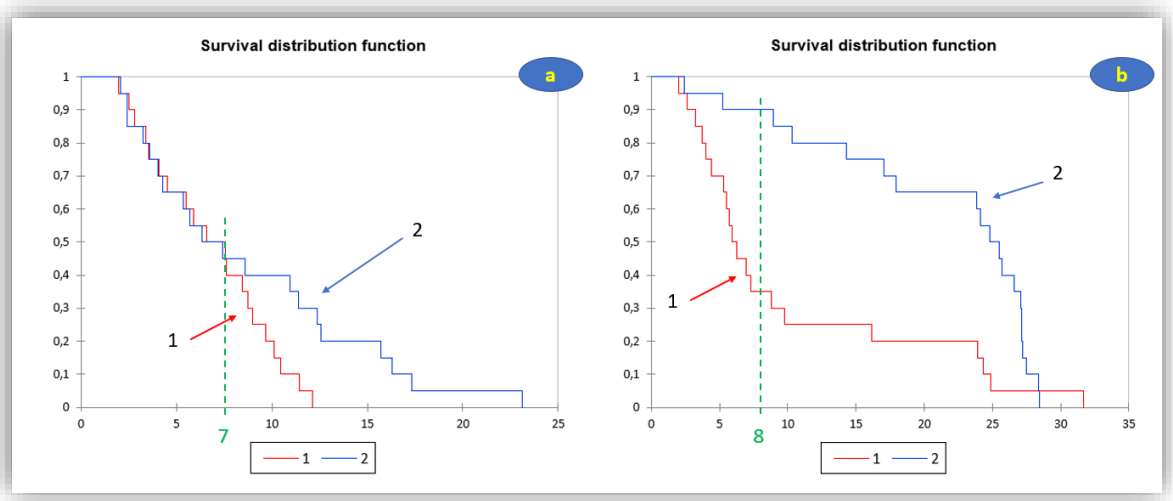


Figure 88

Dans la Figure 88.a, on peut voir que l'événement survient aussi rapidement dans les deux groupes avant t=7 après J0, alors qu'après ce temps-là, l'événement continue de survenir de la même façon au cours du temps dans le groupe 1 alors que dans le groupe 2, la survenue de l'événement se ralentit. Ainsi, l'HRP n'est pas vérifiée pour cette variable relative au groupe car nous sommes alors dans la situation suivante (en supposant que l'on ait le droit de parler de causalité) : « les individus du groupe 1 sont autant à risque d'événement que les individus du groupe 2 peu de temps après J0, mais ils deviennent plus à risque que les individus du groupe 2 plus longtemps après J0 ».

Dans la Figure 88.b, on peut voir que l'événement survient plus rapidement dans le groupe 1 que dans le groupe 2 jusqu'à t=8 après J0, alors qu'après ce temps-là, les choses s'inversent avec une survenue d'événement qui se ralentit dans le groupe 1 alors qu'elle s'accélère dans le groupe 2. Ainsi, l'HRP n'est pas vérifiée pour cette variable relative au groupe car nous sommes alors dans la situation suivante (en supposant que l'on ait le droit de parler de causalité) : « les individus du groupe 1 sont plus à risque d'événement que les individus du groupe 2 peu de temps après J0, mais ils deviennent moins à risque d'événement que les individus du groupe 2 plus longtemps après J0. »

Vous vous rendez bien compte que l'appréciation de l'écart entre deux courbes de Kaplan-Meier est difficile pour apprécier l'écart à l'HRP. D'autres méthodes visuelles sont plus faciles, mais Epi Info ne permet pas de les utiliser.

Ce que vous pouvez vous dire, c'est que l'HRP est vérifiée si les deux courbes de Kaplan-Meier sont soient confondues au cours du temps, soit l'écart entre les deux courbes de survie augmente de plus en plus au cours du temps, et de façon homogène, comme le montre la Figure 89.

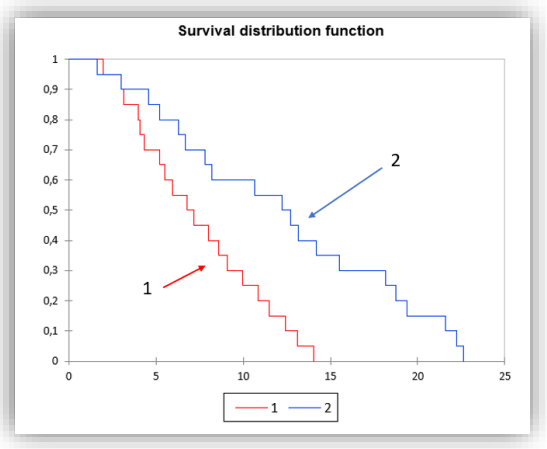


Figure 89

Dans le cas des faibles tailles d'échantillon, si l'HRP est vérifiée au niveau de la population cible, l'écart entre les deux courbes de survie aura malgré tout peu de chances d'augmenter de façon homogène entre les deux courbes estimées à partir des données de l'échantillon. Il deviendra alors difficile de fournir de preuves fortes que l'HRP est vérifiée. Une situation où il est difficile de dire que l'HRP est vérifiée est celle où les deux courbes de survie se croisent à un instant *t* alors qu'avant cet instant *t*, les deux courbes étaient bien séparées (non confondues).

Références

Altman, D.G. and Royston, P., 2006. The cost of dichotomising continuous variables. Bmj. 332, 1080.

Brenner, H. and Blettner, M., 1997. Controlling for continuous confounders in epidemiologic research. Epidemiology. 8, 429-34.

Cox, D.R., 1972. Regression models and life tables (with discussion). J R Statist Soc B. 34, 187-220.

Desquilbet, L. and Mariotti, F., 2010. Dose-response analyses using restricted cubic spline functions in public health research. Stat Med. 29, 1037-57.

Hess, K.R., 1995. Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. Stat Med. 14, 1707-23.

Hua, J., Hoummady, S., Muller, C., Pouchelon, J.L., Blondot, M., Gilbert, C. and Desquilbet, L., 2016. Assessment of frailty in aged dogs. Am J Vet Res. 77, 1357-1365.

Royston, P., Altman, D.G. and Sauerbrei, W., 2006. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 25, 127-41.