

# Les statistiques sous EPIDATA

## Niveau 3 Analysis 2.2

Bernard BRANGER - Réseau « Sécurité Naissance - Naître ensemble » des Pays de la Loire  
1, allée Baco - 44000 NANTES. Tél 02 40 48 55 81 -

Courriel : [bernard.branger@naitre-ensemble-ploire.org](mailto:bernard.branger@naitre-ensemble-ploire.org)

Janvier 2010

### Table des matières

#### Un peu de théorie

<b>I.</b>	<b>Description</b> .....	<b>2</b>
A.	Différentes catégories de variables.....	2
B.	Comment résumer l'information.....	3
1.	Variables quantitatives .....	3
2.	Variables qualitatives .....	5
<b>II.</b>	<b>Comparaisons</b> .....	<b>6</b>
A.	Fluctuations autour d'un pourcentage.....	6
B.	Comparaison de 2 pourcentages.....	8
<b>I.</b>	<b>Variables qualitatives</b> .....	<b>9</b>
A.	Pourcentage observé : freq.....	9
B.	Comparaison de deux pourcentages : tables.....	9
B.	Comparaison de 3 pourcentages ou plus .....	10
<b>II.</b>	<b>Variables quantitatives</b> .....	<b>12</b>
A.	Description d'une moyenne .....	12
B.	Comparaison de deux moyennes .....	15
C.	Comparaison d'une moyenne observée à une moyenne théorique .....	16
D.	Comparaison d'une moyenne à 0.....	16
E.	Comparaison de plusieurs moyennes .....	16
<b>III.</b>	<b>Comparaison de deux variables quantitatives</b> .....	<b>17</b>
<b>IV.</b>	<b>Les différents types d'enquêtes : des indicateurs</b> .....	<b>19</b>
A.	Les enquêtes transversales.....	19
B.	Les enquêtes d'incidence ou enquête de cohorte → RR.....	20
C.	Les enquêtes cas-témoins → OR.....	22
D.	Les études de dépistage ou de diagnostic .....	24
E.	Les courbes de survie .....	24
<b>V.</b>	<b>Pour conclure</b> .....	<b>27</b>
A.	Ouvrages de références.....	27
B.	La partie statistique d'une étude n'est pas la plus importante.....	27
<b>VI.</b>	<b>Tests statistiques usuels</b> .....	<b>28</b>
<b>VII.</b>	<b>Tests et conditions de validité (pages du livre de D. SCHWARTZ)</b> .....	<b>29</b>



## B. Comment résumer l'information

Etablir « des statistiques », c'est dénombrer et résumer l'information pour la transmettre à la communauté médicale. Un résumé oscille toujours entre deux écueils :

« *Tout ce qui est compliqué est incommunicable,  
tout ce qui est simple est faux* »

### 1. Variables quantitatives

Une variable mesurée dans une population prend des valeurs pour les différents individus de cette population : on parle de la DISTRIBUTION des valeurs de la variable. Cette distribution peut être caractérisée par les différents paramètres déjà décrits mais aussi par d'autres se référant à des modèles plus ou moins contraignants.

#### a) Valeur unique qui « résume » les valeurs

→ La variable de l'échantillon est d'allure « normale » ou « gaussienne » ou d'allure symétrique

On utilise une **moyenne**

$$* \text{moyenne} : m \text{ ou } \bar{x} = \frac{\sum x_i}{N} \text{ ou encore } \frac{\sum n_i x_i}{N}$$

				6							
				5	6	7					
			4	5	6	7	8				
	1	2	3	4	5	6	7	8	9	10	11

N=20 ; moyenne = 6

→ La répartition est asymétrique ou les données sont peu nombreuses :

On utilise la **médiane**, valeur telle que 50 % des valeurs soient au-dessus, et 50 % au-dessous. C'est une valeur de rang.

- Nombre de valeurs impaires : 12 15 19 20 23

- idem : 12 15 19 20 56

- Nombre de valeurs paires: 12 15 16 18 20 29 => médiane = 17

				3	4						
			2	3	4						
			2	3	4						
		1	2	3	4	5					
		1	2	3	4	5	6	7			

médiane = 3

#### b) Autres moyens de résumer

→ Etablir des classes

= **Seuils** déterminés selon les habitudes ou les normes internationales :

- l'âge gestationnel avec 37 SA, ou 33 SA, ou 28 SA
- l'IMC à 25 (surpoids) et 30 (obésité) chez l'adulte
- Age des mères tous les 5 ans
- Poids de naissance tous les 500 g : l'habitude veut que le découpage soit de type : 500 g -999 g, puis

1000 g - 1499 g...

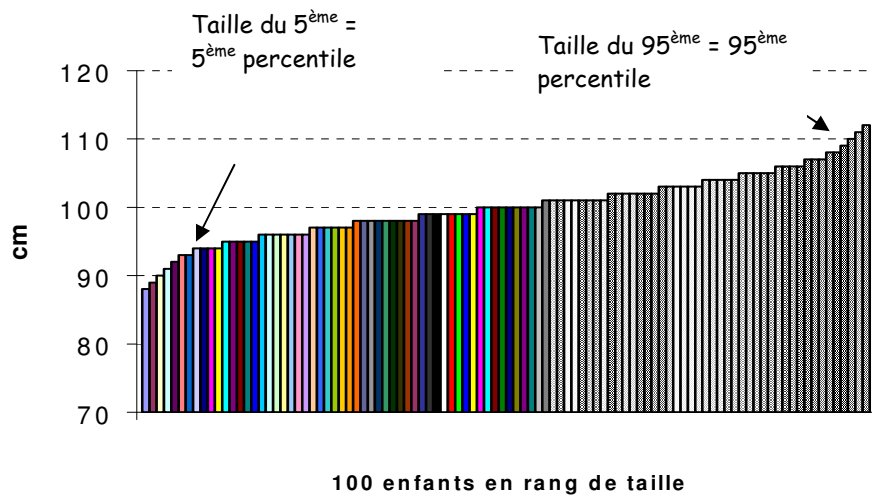
- Sinon, les déterminer soi-même, mais attention aux effectifs et aux risques de données trop personnelles : faire deux classes est toujours un risque d'interprétation ++

= Classes *d'effectifs égaux*

Si l'on découpe l'échantillon en 4 parties égales, on fait apparaître 4 classes séparées par 3 seuils ou QUARTILES : le PREMIER QUARTILE (Q1), valeur pour laquelle 25% des observations sont plus petites et 75% plus grandes. Le DEUXIEME QUARTILE (Q2) n'est autre que la médiane, le TROISIEME QUARTILE (Q3) est la valeur de la variable pour laquelle 75% des observations ont une valeur plus petite et 25% une valeur plus grande.

De la même manière on peut découper la population en TERCILES (par tiers), ou en PERCENTILES : la population est découpée en 100 parties égales. 50<sup>ème</sup> percentile = médiane. La valeur du 3<sup>ème</sup> (ou du 5<sup>ème</sup> percentile) est la valeur correspondant à l'individu qui est au 3% (ou de 5%) de la population ; idem pour le 97<sup>ème</sup> ou le 95<sup>ème</sup>.

Figure 1 : Taille de 100 enfants de 4 ans mis en rang de taille



c) Valeurs de dispersion

Les valeurs centrales comme la moyenne ou la médiane ne suffisent pas : il faut tenir compte de l'étendue (appelée aussi « range » en anglais) des valeurs : le minimum, le maximum.

Lorsque la répartition est symétrique, on peut calculer :

\* *écart-type (déviation standard)*: sorte de « moyenne des écarts à la moyenne »

$$\sigma = \sqrt{\frac{\sum (x_i - m)^2}{N - 1}}, \text{ et la variance : } \sigma^2 = \frac{\sum (x_i - m)^2}{N - 1}$$

Exemple : **population A**. n=20 ; m = 6 ; variance = 122/19 = 6.4 ; écart-type = 2.5

					6						
				5	6	7					
			4	5	6	7	8				
1	2	3	4	5	6	7	8	9	10	11	

Noter que, en cas de distribution dite « normale » au sens de la loi normal, la moyenne et l'écart-type résument à eux seuls les valeurs de la variable.

d) Les liens entre percentiles et écart-type

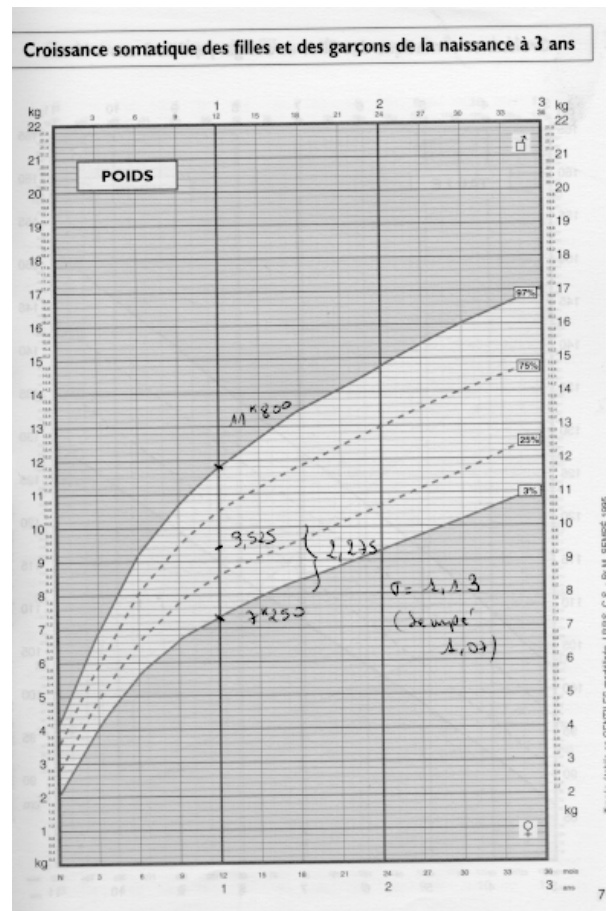
Sous condition d'une distribution normale, 1.96 écarts-type correspond à 2.5 percentiles, ou en arrondissant, - 2 écarts-type au 3<sup>ème</sup> percentile, et + 2 écarts-type au 97<sup>ème</sup> percentile. Lorsque sur une courbe (carnet de santé), on dispose de valeurs, pour un âge donné, du 3<sup>ème</sup> percentile et du 97<sup>ème</sup> percentile, on peut calculer

l'écart-type : il y a 4 écart-types entre ces deux seuils.

Figure 1 bis : Courbe de croissance (carnet de santé)

Sur les carnets de santé les plus récents, les courbes de poids, selon l'âge, sont présentées en percentiles avec la 3<sup>ème</sup>, le 25<sup>ème</sup>, le 75<sup>ème</sup> et le 97<sup>ème</sup>. Curieusement, la médiane (50<sup>ème</sup> percentile) n'est pas mentionnée. L'utilisation de percentiles au lieu des déviations standard utilisées sur les carnets précédents (moyenne et + 2 DS et - 2 DS) est due au caractère asymétrique des poids des enfants de moins de 3 ans : il y a plus de « gros » que de « maigres ».

Sus l'hypothèse (fausse) de répartition symétrique, entre la 97<sup>ème</sup> percentile et le 3<sup>ème</sup>, il y a 2 + 2 écart-types, soit à 1 an,  $11.800 \text{ kg} - 7.250 \text{ kg} = 4.550 \text{ kg}$  ; d'où on calcule que un écart-type =  $4.550 / 4 = 1.13 \text{ kg}$ .



## 2. Variables qualitatives

Une variable qualitative se caractérise par la **FREQUENCE** de ses modalités.

Cette fréquence s'exprime relativement à la population : on a observé 2 fractures de jambe parmi les 20 malades vus aujourd'hui; la fréquence relative est une **PROPORTION** ou un pourcentage de cette pathologie ; dans l'ensemble de la population des malades prise en charge ce jour la proportion est de 10%. Il est indispensable dans l'énoncé d'une fréquence relative de préciser le numérateur ET le dénominateur sous peine d'introduire des ambiguïtés inaccessibles au lecteur. On peut bien entendu regrouper des modalités ou les classes d'une variable qualitative ordinaire ou nominale. Dans la suite de ce document, la fréquence  $f$  sera notée  $p$ , avec  $1 - p = q$ .

On parle également de :

- **TAUX** qui a le plus souvent rapport avec le temps et qui a toujours une unité comme un taux de mortalité, un taux d'incidence et de prévalence.
- **RATIO** qui est un rapport de deux caractéristiques d'une variable, et le numérateur n'est pas contenu dans le dénominateur (et vice versa) et un ratio n'a pas d'unité. Exemple : sex ratio = H/F.
- **INDICE** est un rapport de deux effectifs de nature différente. Exemple : nombre de médecins pour mille habitants.

## II. Comparaisons

Faire une enquête épidémiologique, c'est décrire, mais surtout comparer les malades et les non-malades par exemple pour en déterminer des différences.

**Pourquoi ne pas s'en tenir à l'expression arithmétique d'une proportion : « 30 % », ou « 30 % c'est moins que 40 % » ????**

L'explication des comparaisons entre deux échantillons est difficile à comprendre. Il faut passer par un petit détour.

### A. Fluctuations autour d'un pourcentage

→ **Exemple** : une urne (relativement grande de plusieurs milliers de boules) contient 31 % de boules rouges. On tire au hasard, dans cette urne, 100 boules. Combien de boules rouges peut-on « espérer » ?

**Quel pari peut-on faire ?**

En moyenne : 31 % de 100 = 31... Mais +++++, c'est en moyenne. On trouvera sans doute, si on effectue plusieurs sondages, 29, ou 32, ou 25, ou 38... On pourrait trouver 0 ou 100 boules rouges. **Tout est possible, mais tout n'est pas également probable.**

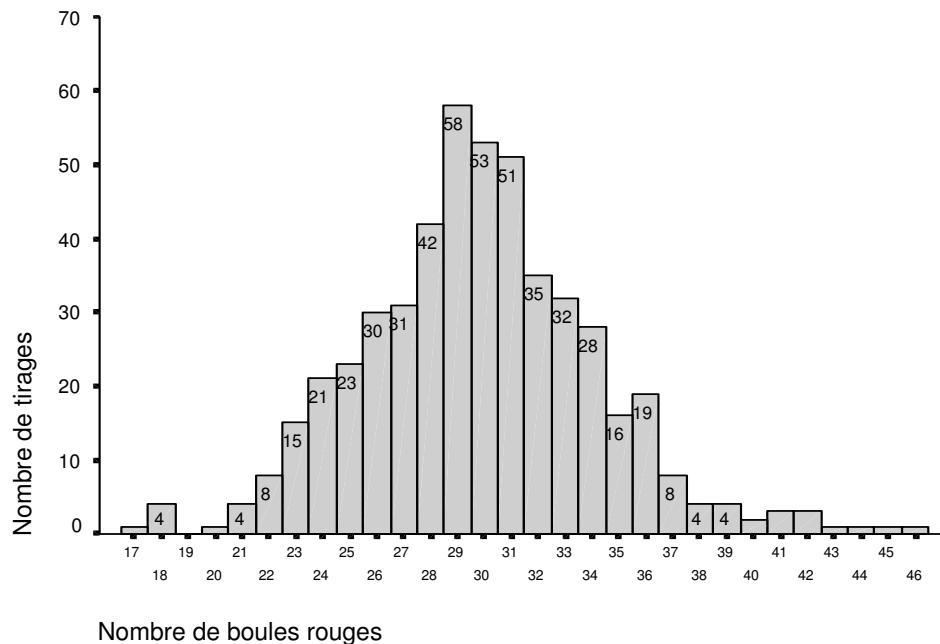
La statistique - corroborant l'observation sur de « grands nombres » - permet de prévoir un intervalle dans lequel on trouvera la majorité des tirages. Cet intervalle, dit « intervalle de pari », est le suivant :

**On a 95 % de chances de trouver entre 22 % et 40 % de boules rouges, soit entre 22 boules et 40 boules.** Autrement dit, à partir de cette urne de 31 % de boules rouges, dans 5 % des cas, on risque de trouver moins de 22 ou plus de 40 boules rouges.

Le calcul est le suivant, en pourcentage (où  $0.69 = 1 - 0.31$ ) ; voir plus loin le pourquoi du « 1.96 »

$$0.31 \pm 1.96 \sqrt{\frac{0.31 * 0.69}{100}} = 0.31 \pm (1.96 * 0.04) = 0.31 \pm 0.09$$

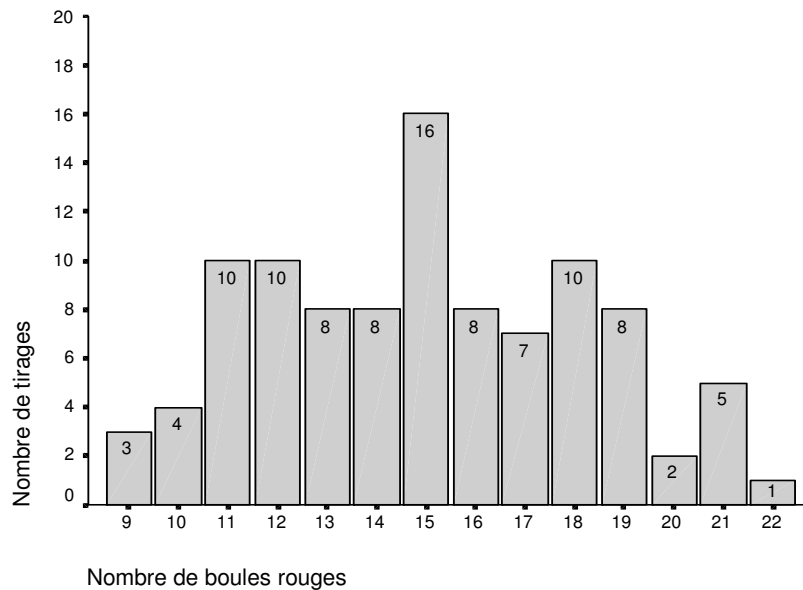
Figure 2 : 500 tirages successifs de 100 boules dans une urne de 31 % de boules rouges



On voit sur ce diagramme que l'on a effectué 500 tirages avec 31 % de boules rouges et 100 boules à chaque fois. La majorité des tirages trouve 29 boules, 30 ou 31...

10 tirages trouvent 21 et moins, et 12 trouvent 39 et plus, soit 22 tirages excessifs, ce qui sur 500, correspond à 4.4 % des tirages.

Figure 3 : Autre démonstration (31 % de 50 boules = 15 boules en moyenne)



100 tirages ont été effectués avec toujours 31 % de boules rouges, mais avec 50 boules chacun : l'intervalle de pari est de 18 % à 44 %, soit 9 à 21 boules autour de 15 boules en moyenne. Cet intervalle est plus large que le précédent (22 % - 40 %), car on tire moins de boules.

D'où le principe à retenir : plus l'échantillon est petit, plus l'intervalle de pari est grand. Le risque de se tromper reste toujours fixé à 5 %, mais l'intervalle est plus large. En d'autres termes, la fourchette entre 22 % et 40 % boules correspond à un intervalle de 95 % avec 100 boules et de 82 % avec 50 boules.

**Si on augmente le nombre de boules, l'intervalle de pari est de plus en plus étroit (au maximum, prendre toutes les boules revient à trouver le pourcentage exact). Si on diminue le nombre de boules, l'intervalle est large, et la prévision est incertaine.**

#### Une autre question : un pourcentage provient-il de l'urne ?

Un autre raisonnement peut alors se proposer : un pourcentage observé  $p_o$  (sur 100 boules) peut-il provenir de l'urne contenant 31 % de boules rouges ? Par rapport à l'intervalle de pari :

→ si  $p_o$  est situé à l'extérieur de l'intervalle 22 % - 40 %, on peut dire que vraisemblablement  $p_o$  ne provient pas de l'urne, mais dans 5 % des cas, on peut se tromper puisque, dans 5 % des cas, l'intervalle de pari est extérieur à la fourchette. On peut même chiffrer le degré d'erreur avec une formule simple et une table :

$$\varepsilon = \frac{p - p_o}{\sqrt{\frac{p * q}{n}}} \quad \text{avec } p = \text{pourcentage théorique et } q = 1-p$$

La lecture dans une table de  $\varepsilon$  permet de déterminer le risque d'erreur (dit  $\alpha$  ou de première espèce) de se tromper quand on conclue à une différence.

→ si  $p_o$  se situe dans la fourchette 22 % - 40 %, on peut dire que  $p_o$  vient de l'urne (en fait, il pourrait s'agir d'une autre urne « équivalente » ou non : on préfère dire que « on ne peut pas montrer que  $p_o$  ne provient pas de cette urne donnée »). Il existe là aussi un risque d'erreur (dit  $\beta$  ou de deuxième espèce) qui consiste à se tromper quand on ne conclue pas à une différence.

= **Exemple** : une proportion de 25 % sur 100 boules (par rapport à l'urne de 31 %) →  $\varepsilon = 1.27$

une proportion de 19 % sur 100 boules (par rapport à l'urne de 31 %) →  $\varepsilon = 2.54$

Si  $\varepsilon > 1.96$ , on dit que  $p_o$  est différent significativement de  $p$ , mais avec un risque d'erreur possible, visible dans la table (exemple :  $\varepsilon = 2.57$ , risque de 0.01 ; ou  $\varepsilon = 3.29$ , risque de 0.001). Si  $\varepsilon \leq 1.96$ , on dit que  $p_o$  n'est pas différent de  $p$  de l'urne théorique.

**Un pourcentage observé est assorti d'un « intervalle de confiance »**

Maintenant, on se situe en clinique habituelle : on observe un  $p_o$  sur  $n$  sujets. Quel est la vraie proportion, puisque une fluctuation due au hasard existe toujours ? C'est l'intervalle de confiance qui nécessite de connaître  $p_o$  et  $n$ . La formule est la suivante pour un intervalle de confiance à 95 % (ou en d'autres termes : on a 95 % de chance que la vraie proportion se situe entre les deux bornes :

$$p \text{ théorique} = p_o \pm 1.96 \sqrt{\frac{p_o * q_o}{n}}$$

Exemple : une proportion de 12 % sur 156 sujets :  $0.12 \pm 0.05$ , soit de 0.07 à 0.17.

**Au total**

- = En épidémiologie, on travaille toujours sur des ECHANTILLONS
  - = La science est toujours GÉNÉRALE et la technique statistique consiste à essayer de passer de l'échantillon à la POPULATION : à généraliser pour faire progresser la Science
  - = On connaît souvent bien les numérateurs... toujours bien connaître, et bien situer les dénominateurs +++
  - = Toujours dire sur combien de sujets les pourcentages sont calculés
  - = Donner l'intervalle de confiance à 95 % des pourcentages observés (I.C. à 95 %)
  - = Est-ce qu'il y a un IC à 95 % correct ? Plus l'échantillon est grand, plus l'intervalle est petit....
- NB : il faut suffisamment de sujets pour appliquer la formule. Voir infra.*

**B. Comparaison de 2 pourcentages**

Le raisonnement est le même : si la différence entre les deux pourcentages est grande, on peut dire que probablement ils sont issus de deux urnes différentes (au risque d'erreur près que l'on peut chiffrer). Si la différence est petite, on ne peut pas montrer que les deux pourcentages sont différents. Le principe est un calcul de  $\varepsilon$  :

$$\varepsilon = \frac{p_A - p_B}{\sqrt{\frac{p * q}{n_A} + \frac{p * q}{n_B}}} \text{ avec } p = \frac{n_A p_A + n_B p_B}{n_A + n_B} \text{ et } q=1-p$$

Si  $\varepsilon > 1.96$ , on conclue que les deux pourcentages sont différents (avec un risque d'erreur de type  $p < 0.05$ ).  
Si  $\varepsilon \leq 1.96$ , on ne conclue pas, ou on conclue qu'ils ne sont pas différents.

La plupart des logiciels et des livres de statistiques utilisent plutôt le  $\chi^2$  (chi-carré ou chi-deux) qui calcule les écarts entre les effectifs observés et les effectifs calculés sous l'hypothèse d'égalité des deux pourcentages. La valeur limite du  $\chi^2$  pour deux pourcentages est de 3.84 (la carré de 1.96) pou un degré de signification de 0.05 . Voir infra.

Pour les savants, la formule du  $\chi^2$  est la suivante :

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{o_i} \text{ ou encore } \chi^2 = \frac{N(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

## Commandes EPIDATA dans Analysis 2.2 - Comment interpréter

Une des étapes fondamentales est de connaître les formes des variables :

= **variable qualitative** (oui/non, codes 1-2-3...) : description et comparaison de pourcentages (varqual)

= **variable quantitative** (valeurs numériques, forme en #### dans les \*.qes) : description et comparaison de moyennes (varquant)

Les exemples sont tirés de deux enquêtes : l'une concernant l'allaitement maternel sur 239 femmes de 18 maternités (enquête de cohorte), l'autre concernant des nouveau-nés macrosomes comparés à des témoins (enquête cas-témoins).

*NB : Pour la clarté, les tableaux comportent lignes horizontales et verticales, ce qui n'est pas le cas en sortie de EPIDATA.*

### I. Variables qualitatives

#### A. Pourcentage observé : freq

La commande **freq varqual** donne la fréquence (avec /c) et l'intervalle de confiance à 95 % (avec /ci).

Exemple : % de femmes ayant préparé l'accouchement oui / non

→ Exemple : **freq prepa /c** pour les pourcentages en colonne

Prepa		
	N	%
1	175	73.2
2	64	26.8
Total	239	100.0

→ Exemple : **freq prepa /c /ci** pour les pourcentages en colonne et l'intervalle de confiance à 95 %.

Prepa			
	N	%	(95% CI)
1	175	73.2	(67.3-78.4)
2	64	26.8	(21.6-32.7)
Total	239	100.0	

→ Exemple : **freq activmere /c /cum** pour les pourcentages en colonne et les pourcentages cumulatifs

Activ mere			
	N	%	Cum %
1	1	0.4	0.4
2	6	2.5	3.0
3	43	18.1	21.1
4	30	12.7	33.8
5	103	43.5	77.2
6	9	3.8	81.0
7	23	9.7	90.7
8	22	9.3	100.0
Total	237	100.0	

NB : on peut noter à la suite : **freq var1 var2 var3** pour obtenir 2 tableaux successifs selon var3

#### B. Comparaison de deux pourcentages : tables

La commande **tables varqual1 varqual2** permet de comparer deux pourcentages ou plus. En ajoutant /r, on obtient les % en lignes, ou /c en colonnes. L'ajout de /t donne le test du  $\chi^2$ , de /ex le test de Fisher.

Le test statistique permet aussi de vérifier que l'on a le droit de conclure à une différence compte tenu des fluctuations d'échantillonnage. Le test statistique est le  $\chi^2$  dont la signification tourne autour de 3.84 pour

être significatif (pour deux pourcentages). En cas d'effectifs insuffisants (effectifs calculés < 5), ne pas conclure avec le  $\chi^2$ . Exemple : croisement de préparation à l'accouchement et de la catégorie socioprofessionnelle du couple (cadre ou non).

→ **Exemple : tables prepa couple** pour croiser deux variables ici codées en 2 classes chacune (avec « couple » codé 3 = profession cadre d'un des membres du couple, et 8 = pas cadre. « Prepa » = Préparation à l'accouchement codée 1 pour oui et 2 pour non.

Prepa			
COUPLE	1	2	Total
3	82	19	101
8	93	45	138
Total	175	64	239

→ **Exemple : tables prepa couple /r /t** pour les pourcentages en lignes et le test du  $\chi^2$  (significatif)

Prepa						
COUPLE	1	%	2	%	Total	%
3	82	(81.2)	19	(18.8)	101	(100.0)
8	93	(67.4)	45	(32.6)	138	(100.0)
Total	175	(73.2)	64	(26.8)	239	
Percents: (Row) Chi2= 5.662 df(1) p= 0.0173						

Le test du  $\chi^2$  compare 81.2 % (proportion de préparation quand le couple est classé « cadre ») à 67.4 % (proportion de préparation quand le couple est classé « pas cadre ») : cette différence est significative avec un risque de se tromper de 1.73 %. On dit que la proportion de préparation est significativement différente entre les cadres et les non-cadres.

→ **Exemple : tables prepa couple /r /t /ex** avec le test de Fisher en cas de petits effectifs

Prepa						
COUPLE	1	%	2	%	Total	%
3	6	(100.0)	0	(0.0)	6	(100.0)
8	11	(68.8)	5	(31.3)	16	(100.0)
Total	17	(77.3)	5	(22.7)	22	
Percents: (Row) Chi2= 2.426 df(1) p= 0.1193						
Cells expected*5: 3 (75 pct.)						
Fisher's exact p= 0.2663						

Ce tableau permet de comparer une proportion de 100 % à 68.8 % avec des effectifs faibles. Ce n'est pas la cellule avec « 0 » qui est gênante mais les effectifs calculés sous l'hypothèse d'égalité des %...

C'est le test de Fisher qu'il faut demander et interpréter (NS). Remarque : df = degree of freedom = degré de liberté en français = 1 en cas de deux pourcentages.

## B. Comparaison de 3 pourcentages ou plus

C'est la même commande **tables varqual1 varqual2** où un des modalités est codée en 3 classes ou plus. Les conditions d'application sont les mêmes : tous les effectifs doivent être > 5. Lorsqu'un effectif < 5, le test de Fisher n'est pas valable pour plus de 2 pourcentages. Le test de Yates non plus. Il faut alors regrouper les classes. Tout regroupement est une interprétation qui peut être critiquable.

→ **Notion de degrés de liberté** (ddl, ou degree of freedom, df, en anglais). La valeur du  $\chi^2$  dépend du nombre de pourcentages à comparer. On le calcule avec la formule  $ddl = (c-1)(l-1)$  où c est le nombre de colonnes et l le nombre de lignes. Soit pour 3 pourcentages : ddl = 2 (pour 2 pourcentages, c'est 1). La valeur limite du  $\chi^2$  pour 2 ddl est de 5.99.

→ Exemple : tables prepa episio /t

Episio	Prépa(ration à l'accouchement)				Total	%
	1	%	2	%		
1 Oui	37	(82.2)	8	(17.8)	45	(100.0)
2 Non	61	(71.8)	24	(28.2)	85	(100.0)
3 Césarienne	14	(77.8)	4	(22.2)	18	(100.0)
Total	112	(75.7)	36	(24.3)	148	

Percents: (Row) Chi2= 1.797 df(2)  
p= 0.4071  
Cells expected<5: 1 (17 pct.)

On ne peut interpréter ce tableau car le  $\chi^2$  n'est pas valable et le Fisher ne vaut que pour deux pourcentages à comparer. Il faut regrouper ou ne sélectionner que deux lignes : ici, le plus évident est de ne pas tenir compte de la césarienne.

→ Exemple

select episiotomie<3  
tables prepa episio /t

Episiotomie	Préparation à l'accouchement				Total	%
	1	%	2	%		
Oui	37	(82.2)	8	(17.8)	45	(100.0)
Non	61	(71.8)	24	(28.2)	85	(100.0)
Total	98	(75.4)	32	(24.6)	130	

Percents: (Row) Chi2= 1.734 df(1) p= 0.1879

#### 4. Comparaison de deux variables en tenant compte d'une troisième (ajustement)

Le lien entre arrêt de l'allaitement à 1 mois est lié au moment de la décision d'allaiter : lorsque la décision d'allaiter a été prise avant la grossesse, l'arrêt à 1 mois est de 6.0 % versus 19.6 % lorsque la décision a été prise pendant la grossesse (p=0.0025) : voir tableau ci-dessous avec tables arretm1 moment /r /t /ex

arretm1						
Moment	1	%	2	%	Total	%
1	10	(6.0)	157	(94.0)	167	(100.0)
2	11	(19.6)	45	(80.4)	56	(100.0)
Total	21	(9.4)	202	(90.6)	223	

Unstratified table Chi2= 9.167 df(1) p= 0.0025  
Fisher's exact p= 0.0061

Ce lien persiste-t-il lorsque l'on tient compte de la primiparité (versus multiparité) ? On réalise alors deux tableaux en croisant moment avec arrêt : l'un pour les primipares, l'autre pour les multipares avec une seule commande :

→ Exemple : tables arretm1 moment primi /r /t /ex

primi: 1						
arretm1						
Moment	1	%	2	%	Total	%
1	3	(5.3)	54	(94.7)	57	(100.0)
2	10	(28.6)	25	(71.4)	35	(100.0)
Total	13	(14.1)	79	(85.9)	92	

Percents: (Row) Chi2= 9.709 df(1) p= 0.0018  
Cells expected<5: 1 (25 pct.)  
Fisher's exact p= 0.0038

<b>primi: 2</b>						
arretm1						
Moment	1	%	2	%	Total	%
1	7	(6.4)	103	(93.6)	110	(100.0)
2	1	(4.8)	20	(95.2)	21	(100.0)
Total	8	(6.1)	123	(93.9)	131	
Percents: (Row) Chi2= 0.079 df(1) p= 0.7788 Cellsexpected<5: 1 (25 pct.)Fishersexact p= 1.0000						

Le test est significatif pour les primipares, mais pas pour les multipares.

Le test global dit de Mantel-Haenszel est significatif (même commande ; résultat à la suite) : on dit que l'ajustement sur la parité ne modifie pas de manière globale le lien entre arrêt et moment, mais c'est surtout valable pour les primipares.

<b>arretm1 by Moment adjusted for primi</b>								
N = 223	N	Chi2	Df	p	p exact			
Crude	223	9.167	1	0.002	0.0061			
<b>Adjusted</b>	<b>223</b>	<b>6.19</b>	<b>1</b>	<b>0.013</b>	<b>0.0100</b>			
primi: 1	92	9.709*	1	0.002	0.0038			
primi: 2	131	0.079*	1	0.779	1.0000			
Summary Estimates								
Total 2 strata. 2 informative & 0 non-informative.								
*: Small Expected Numbers, use P exact /ex Exact								
Monte Carlo tests: 500 simulations Chi2 = 9.788 df(1) p=0.0100								

## II. Variables quantitatives

### A. Description d'une moyenne

EPI-DATA donne automatiquement les caractéristiques d'une variable quantitative.

La commande freq n'est pas utile : elle liste tous les âges.

→ Exemple : **freq age /c /cum** pour avoir les proportions par âge et les proportions cumulées

Age	N	%	Cum %	Age	N	%	Cum %
16	1	0.7	0.7	32	9	6.1	61.2
17	1	0.7	1.4	33	8	5.4	66.7
20	1	0.7	2.0	34	10	6.8	73.5
21	2	1.4	3.4	35	9	6.1	79.6
23	4	2.7	6.1	36	3	2.0	81.6
24	1	0.7	6.8	37	6	4.1	85.7
25	5	3.4	10.2	38	8	5.4	91.2
26	10	6.8	17.0	39	6	4.1	95.2
27	10	6.8	23.8	40	3	2.0	97.3
28	14	9.5	33.3	41	1	0.7	98.0
29	10	6.8	40.1	42	1	0.7	98.6
30	9	6.1	46.3	44	1	0.7	99.3
31	13	8.8	55.1	45	1	0.7	100.0
<i>suite colonne en haut</i>				<b>Total</b>	<b>147</b>	<b>100.0</b>	

→ Exemple : means age pour l'âge des mères (en années)

Age	Obs.	Sum	Mean	Variance	Std Dev	( 95% CI mean )		Std Err	
	147	4588.00	<b>31.21</b>	27.24	<b>5.22</b>	30.36	32.06	0.43	
	Minimum	p5	p10	p25	Median	p75	p90	p95	Max
	<b>16.00</b>	23.00	25.00	28.00	<b>31.00</b>	35.00	38.00	39.60	<b>45.00</b>

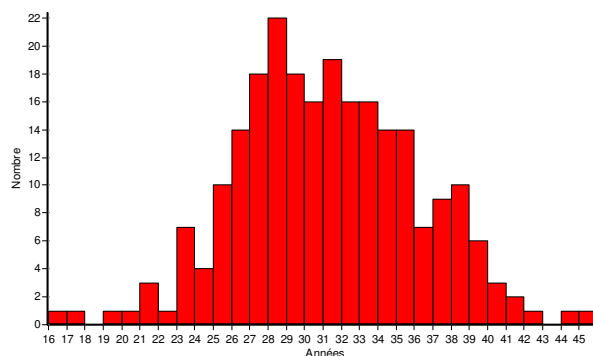
- **Mean** = moyenne est la division de la somme des âges (sum) par le nombre de mères (obs.) :  $4588 / 147 = 31.21$ . A noter que ce sont des décimales d'années (et non des mois +++): ici 0.21 d'année = 2.5 mois environ.
- **Variance** vaut 27.24 et est le carré de l'écart-type
- **Std Dev** = **Ecart-type** pour déviation standard est la racine carrée de la variance : rac. car. de 27.24 = 5.22 années (c'est donc la moyenne des écarts à la moyenne). Il reflète la dispersion des valeurs.
- **Std Err** = **erreur-type** est l'écart-type de la moyenne, soit écart-type / racine de 147 = 0.43 (pas utile en pratique)

- La ligne suivante est une description des rangs des valeurs avec le **minimum** et le **maximum**
  - = le **quartile 25 (Q1)** : la valeur du 25<sup>ème</sup> percentile (environ au niveau de 25 % dans les pourcentages cumulatifs)
  - = le **quartile 75 (Q3)** : la valeur du 75<sup>ème</sup> percentile (environ au niveau de 75 % dans les pourcentages cumulatifs)
  - = la **médiane** est la valeur du 50<sup>ème</sup> percentile (ou Q2) (50 % en cumulatif). Ici, la médiane est pratiquement égale à la moyenne : preuve d'une symétrie de la répartition. Les médianes sont du côté des pics et les moyennes sont attirées par les extrêmes.

→ Exemple : histogram age /edit pour obtenir des barres verticales contiguës année / année

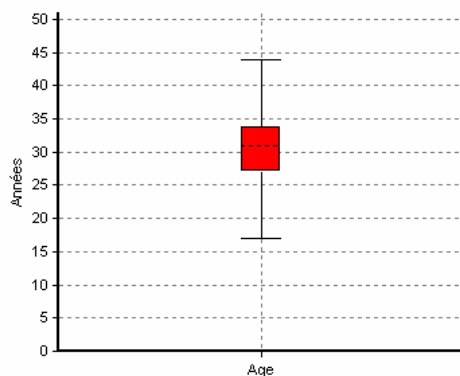
Si les âges sont avec des décimales, le module Graph d'Epidata ne regroupe pas en âge entier il faut créer une nouvelle variable (exemple : define agent ## puis ageint = integer(age) puis histogram agent)

Figure 4 : Histogramme des âges



→ Exemple : box Age /edit pour obtenir un graphe dit en box-plot

Figure 4 bis : Box-plot des âges



Le trait en pointillé au milieu de la boîte est la médiane.

Le trait du bas de la boîte est le 1<sup>er</sup> quartile ou le 25<sup>ème</sup> percentile (Q1), et celui du haut le 3<sup>ème</sup> quartile ou le 75<sup>ème</sup> percentile (Q3). Chaque trait au bout de la ligne correspond à médiane  $\pm 1.5 \cdot (Q3 - Q1)$ ; des « outliers » ou valeurs extrêmes peuvent apparaître au-delà de ces deux limites.

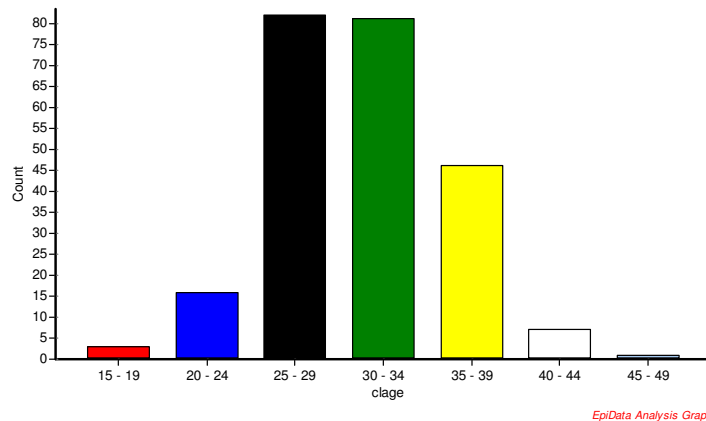
- **Exemples** : comment transformer une variable quantitative en classes avec seuils réguliers
  - define classage #** pour créer une nouvelle variable (ici numérique)
  - recode age to classage by 5** (classe tous les 5 ans)
  - freq classage /c** (pour la fréquence)

**Age en classe de 5 ans**

Classe	N	%
15 - 19	2	1.4
20 - 24	8	5.4
25 - 29	49	33.3
30 - 34	49	33.3
35 - 39	32	21.8
40 - 44	6	4.1
45 - 49	1	0.7
Total	147	100.0

- **Exemple** : **bar classage /edit** pour obtenir un graphe en bâtons (non contiguës). Les couleurs sont à changer (un peu long, préférer le Graph de Word ?).

**Figure 5 : Diagramme en bâtons (ou barres) de classe d'âges**



**Notes personnelles**

## B. Comparaison de deux moyennes

Exemple de commande générale : **means age moment** avec la variable quantitative d'abord puis la variable qualitative. Lire les résultats en face de la mention du test de t. Le test de EPIDATA est une analyse de variance (ANOVA), mauvais mot car c'est une analyse de moyenne +++

→ Quelques formules à ne pas retenir ++

= Les deux échantillons ont plus de 30 sujets chacun : la comparaison de deux moyennes utilise le test  $\epsilon$  avec la limite de 1.96. La formule est la suivante :

$$\epsilon = \frac{m_A - m_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \text{ où } \sigma_A^2 \text{ et } \sigma_B^2 \text{ sont les variances dans chaque échantillon A et B}$$

= Au moins un des échantillons a moins de 30 sujets et la répartition est normale, c'est le test t de Student qui est indiqué :

$$t_{ddl} = \frac{m_A - m_B}{\sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B}}} \text{ où } \sigma^2 \text{ est la variance commune (calcul hors programme)}$$

La valeur de t se lit dans une table en fonction du nombre de degrés de liberté (ddl) avec  $ddl = n_A + n_B - 2$ . Quand le test de Student est utilisé pour  $n > 30$ , il vaut 1.96. D'autres conditions sont nécessaires : que les répartitions soient normales (voir précédemment et infra) et que les variances soient semblables.

= Les échantillons sont petits et il y a un doute sur la normalité de la répartition, un test non-paramétrique est nécessaire (test de Wilcoxon ou Mann-Whitney pour deux moyennes, de Kruskal-Wallis pour deux moyennes ou plus).

→ Exemple : **means age /by=prepa /t** pour deux moyennes d'âge de prepa et pour le « p »

Age	Obs.	Sum	Mean	Variance	Std Dev	( 95% CI mean )	Std Err
1	112	3476.00	<b>31.04</b>	24.72	4.97	30.10 31.97	0.47
2	35	1112.00	<b>31.77</b>	35.83	5.99	29.72 33.83	1.01

Prepa	Minimum	p5	p10	p25	Median	p75	p90	p95	Max
1	17.00	23.00	25.00	28.00	31.00	34.00	38.00	39.35	45.00
2	16.00	20.00	24.80	28.00	32.00	36.00	39.00	42.40	44.00

Source	SS	df	MS	F	p Value
Between	14.43	1	14.43	0.53	<b>0.469</b>
Within	3962.03	145	27.32		
Total	3976.46	146	27.24		

Bartlett's test for homogeneity of variance → Chi2= 1.884 df(1) p= 0.170

Il faut lire les données suivantes :

- Les deux moyennes sous « mean » 31.0 et 31.8 (en arrondissant) : ce sont elles que l'on compare
- Le p value situé à droite à 0.469 : il est  $> 0.05$  → les différences ne sont pas significatives.
- Le p du bas à 0.170 veut dire que les variances ne sont pas différentes (test de Bartlette, dans d'autres logiciels test de Levène).

→ Exemple **kwallis age /by=prepa** pour utiliser le test de Kruskal-Wallis en cas de petits effectifs

Kruskal - Wallis One-Way Analysis of Variance		
Ranks of Age		
Prepa	N	Sum of Ranks
1	112	8100.00
2	35	2778.00
Chi2 = 0.7311 df( 1) p= 0.3925 - 18 ties		
Corrected for Ties: Chi2 = 0.7340 Df( 1) p= 0.3916		

Le "p" à retenir est ici 0.39 (NS).

### C. Comparaison d'une moyenne observée à une moyenne théorique

Il n'y a pas de test dans EPIDATA. A la main, le test est le suivant :

$$\varepsilon = \frac{m_o - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

avec  $m_o$  = moyenne observée, et  $\sigma$  = écart-type de l'échantillon

La valeur de  $\varepsilon$  se lit par rapport à 1.96 (voir supra)

### D. Comparaison d'une moyenne à 0

On peut être amenée à faire une différence de deux moyennes pour vérifier si cette différence « diffère » de 0 : le test de comparaison à 0 se fait de la manière suivante par exemple pour explorer la différence entre la pH artériel et le pH veineux :

#### Exemples

```
define diff ##.##
diff = pha - phv
means diff /t
```

diff	Obs.	Sum	Mean	Variance	Std Dev	( 95% CI mean )		Std Err	
	128	-10.36	-0.0809	0.0026	0.0513	-0.0899	-0.0720	0.00	
	Minimum	p5	p10	p25	Median	p75	p90	p95	Max
	-0.240	-0.150	-0.140	-0.110	-0.0800	-0.0500	-0.0200	-0.0100	0.140
Students T-test for mean=0: T= 17.86 df(127) p=0.00000									

Le test de Student à lire est dans la dernière ligne : la moyenne  $-0.0809 \pm 0.05193$  est différente de 0 avec  $p < 10^{-5}$ .

### E. Comparaison de plusieurs moyennes

C'est le test ANOVA dont l'explication sort de ce cours. La présentation des données : la variable quantitative avec une ligne par sujet et une variable qualitative de groupe. Conditions : la population doit être de répartition normale.

→ Exemple : means age /by=episio /t pour comparer l'âge des mères selon la pratique d'une épisiotomie (code en 3 classes : oui, non, césarienne)

Episio	Obs.	Sum	Mean	Variance	Std Dev	( 95% CI mean )		Std Err	
1	45	1340.00	<b>29.78</b>	24.77	4.98	28.28	31.27	0.74	
2	84	2687.00	<b>31.99</b>	26.64	5.16	30.87	33.11	0.56	
3	18	561.00	<b>31.17</b>	31.32	5.60	28.38	33.95	1.32	

Episio	Minimum	p5	p10	p25	Median	p75	p90	p95	Max
1	16.00	21.60	23.60	26.50	29.00	33.50	37.40	38.70	40.00
2	17.00	23.50	26.00	28.00	32.00	35.00	39.00	40.00	45.00
3	20.00	20.00	25.00	27.00	31.00	35.00	39.00	39.00	42.00

Source	SS	df	MS	F	p Value
Between	143.20	2	71.60	2.69	<b>0.0713</b>
Within	3833.27	144	26.62		
Total	3976.46	146	27.24		
Bartlett's test for homogeneity of variance Chi2= 0.344 df(2) p= 0.842					

Il n'y a pas de différence d'âges entre les trois groupes ( $p=0.07$ ). Remarque : la troisième codage n'est pas très adapté : il faut sans doute n'étudier le lien que pour les voies basses.

→ Exemple :

```
select episio <> 3 ou idem select episio <3 ou idem select (episio=1) or (episio = 2)
means age /by=episio /t
```

Age

Episio	Obs.	Sum	Mean	Variance	Std Dev	( 95% CI	mean )	Std Err	
1	45	1340.00	<b>29.78</b>	24.77	4.98	28.28	31.27	0.74	
2	84	2687.00	<b>31.99</b>	26.64	5.16	30.87	33.11	0.56	

Episio	Minimum	p5	p10	p25	Median	p75	p90	p95	Max
1	16.00	21.60	23.60	26.50	29.00	33.50	37.40	38.70	40.00
2	17.00	23.50	26.00	28.00	32.00	35.00	39.00	40.00	45.00
Source	SS	df	MS	F	p Value				

Between	143.16	1	143.16	5.51	<b>0.0205</b>
Within	3300.77	127	25.99		
Total	3443.92	128	26.91		
Bartlett's test for homogeneity of variance Chi2= 0.075 df(1) p= 0.784					

Analyse : l'âge est différent avec p = 0.02 entre les femmes avec épisiotomie et les autres sans.

### III. Comparaison de deux variables quantitatives

Lorsque l'on veut étudier si deux variables quantitatives sont liées, on peut d'abord dresser un graphe avec l'une en abscisse et l'une en ordonnée. C'est généralement un « scatter plot » ou diagramme en points

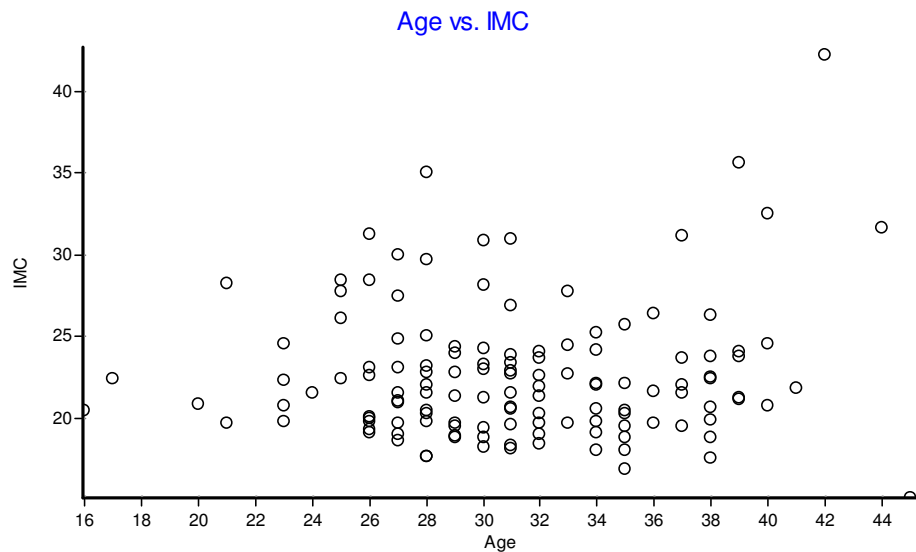
→ Exemple : scatter varquant1 varquant2 (ou scatter x y)

Chaque point correspond à un couple de valeurs : age et IMC.

De même qu'une moyenne résume un donnée, les couples de données peuvent être « résumer » par une droite : c'est la « droite de régression » . EPI-INFO en donne la formule de type  $y = A \cdot X + B$  où B est l'ordonnée à l'origine et A la pente de la droite (avec une unité).

Exemple : scatter age imc

Figure 6 : Scatter plot (diagramme en points)



EpiData Analysis Graph

A signaler que EPIDATA, malheureusement pour l'instant, ne peut pas présenter la droite de régression (demande auprès du manager)...

→ **Exemple : regress varquant1 varquant2** (ou regress y x)

L'hypothèse de départ est qu'il n'y a pas de lien entre les deux variables et que la droite est horizontale (quelle que soit les valeurs de x, y a toujours la même valeur), ou encore que la pente est nulle. Le test dit « test de la pente » consiste à montrer que la pente est différente de 0. Un  $p < 0.05$  signifie que la pente est significativement différente de 0 et que les variables sont liées entre elles.

**Exemple : regress imc age**

Source	Sum Sq	Mean Sq	df		Number of obs	131
Model	16.05	16.05	1		F(1,129)	0.92
Residual	2246.07	17.41	129		Prob > F	<b>0.34</b>
Total	2262.11	17.40	130		<b>R-squared</b>	<b>0.01</b>
					Root MSE	4.17
Variable	Beta	LCL	UCL	SE	t	P> t
<b>age</b>	<b>0.07</b>	-0.07	0.20	0.07	0.96	<b>0.34</b>
<b>Intercept</b>	<b>20.58</b>	16.32	24.85	2.16	9.55	0.00

Total N = 150 Included: N= 131

La droite de régression a la formule suivante :  $imc = 0.07 * age + 20.58$

La pente A vaut 0.07 (donc croissante et son « p » / à l'hypothèse de pente nulle vaut .034 donc NS).

**Pour mieux comprendre : le  $r^2$  (R-squared)**

Un  $r^2$  peut s'interpréter comme la proportion de la variance (de la variabilité) qui est expliquée par le facteur étudié. Ici 0.01, soit 1 % de l'IMC est expliqué par l'âge...

→ **Exemple : correlate age imc**

Cette commande permet d'établir un coefficient de corrélation dit « r » : il va de -1 à +1 en passant par 0. Plus il est prêt de -1 ou +1, plus les deux variables sont corrélées, soit positivement (quand l'une augmente, l'autre augmente), soit négativement (quand l'une augmente, l'autre diminue). Ici, il vaut 0.084 : son « p » est le même que celui de la pente précédente ( $p = 0.34$ , NS).

age	imc	
age	1.000	
imc	<b>0.084</b>	1.000

Total N = 150 Included: N= 131

A signaler le **coefficient de corrélation  $\rho$  de Spearman** qui est un r sur les données ordonnées (classées en rang) : pour les petits effectifs sans hypothèse de distribution (dans SPSS).

## IV. Les différents types d'enquêtes : des indicateurs

### A. Les enquêtes transversales

Les enquêtes de prévalence n'ont pas d'interprétation spécifique. Ce sont plutôt des enquêtes de description. **Les comparaisons** (pourcentages ou variables quantitatives) **sont souvent hasardeuses** car les populations sont trop différentes. En particulier, les infections nosocomiales sont issues d'une population de malades encore traités pour une infection, tandis que les non-infections sont plus aléatoires... Il est normal, par exemple, que les durées de séjour (au jour de l'enquête) soient différentes puisque les définitions ne sont pas les mêmes. C'est pour cette raison que les comparaisons sont délicates.

Il n'est pas d'usage d'utiliser des risques relatifs ou des odds ratios en enquête de prévalence (certains parelent de rapport de prévalence avec des « faux » RR (voir infra).

Exemple : Enquête sur l'utilisation de médicaments tocolytiques en cas de MAP sur placenta prævia ? Certaines femmes ont eu 1, 2 ou 3 produits :

Exemple : freq betabloq adalate tractocile /c

BETABLOQ			ADALATE			TRACTOCILE		
	N	%		N	%		N	%
1 (oui)	102	34.9	1	59	20.2	1	30	10.3
2 (non)	190	65.1	2	233	79.8	2	262	89.7
Total	292	100.0	Total	292	100.0	Total	292	100.0

Exemple de transformation en 3 classes (sans le tractocile à part...)

#### → Exemples

```
define classe #
if (betabloq=1) then classe=1
if (adalate=1) then classe=2
if (betabloq=1) and (adalate=1) then classe=3
if classe=. then classe=4
freq classe /c
```

CLASSE			
	N	%	
1	83	31.7	pour le seul betabloquant
2	29	11.1	pour le seul adalate
3	11	4.2	pour betabloquant + adalate
4	139	53.1	pour l'absence de traitement
Total	262	100.0	

On désire étudier le terme de naissance selon le produit

Exemple : means ag classe /t

AG									
CLASSE	Obs.	Sum	Mean	Variance	Std Dev	( 95% CI	mean )	Std Err	
1	83	2778.14	<b>33.47</b>	13.88	3.73	32.66	34.29	0.41	
2	29	995.02	<b>34.31</b>	8.44	2.91	33.21	35.42	0.54	
3	11	383.28	<b>34.84</b>	8.57	2.93	32.88	36.81	0.88	
4	139	5151.7	<b>37.06</b>	12.13	3.48	36.48	37.65	0.30	
CLASSE	Minimum	p5	p10	p25	Median	p75	p90	p95	Max
1	25.71	27.34	28.63	30.29	33.43	36.29	38.03	39.43	41.29
2	27.57	28.22	30.71	32.86	34.00	36.36	37.43	39.79	41.29
3	30.00	30.00	30.00	32.71	34.14	37.86	39.71	39.71	39.71
4	25.00	28.43	31.86	36.00	37.86	39.14	40.43	41.14	41.71
Source	SS	df	MS	F	p Value				
Between	726.50	3	242.17	19.93	<b>0.00000000</b>				
Within	3134.77	258	12.15						
Total	3861.27	261	14.79						
Bartlett's test for homogeneity of variance Chi2= 2.909 df(3) p= 0.406									

Il y a de grandes différences de terme selon la classe de produit : quand on ne donne rien, le terme est supérieur de 4 semaines / betabloquant et de 3 semaines / adalate.... Attention au risque de raisonnement inversé ; ce n'est pas parce que l'on a donné de l'adalate que la naissance a eu lieu à 34 SA, mais parce que la MAP était importante que l'on a donné de l'adalate ++++

### B. Les enquêtes d'incidence ou enquête de cohorte → RR

Ce type d'enquête permet de suivre un groupe de personnes exposées à un facteur et qui deviendront malades ou non. On pourra calculer et les fréquences de l'exposition et la fréquence (incidence) de la maladie. Les comparaisons sont licites et font partie des objectifs de ce genre d'enquêtes. Outre les comparaisons de pourcentages ou de moyennes, un indice particulier est à utiliser : **le risque relatif (RR)**.

→ Exemple : tables arretm1 moment /r /t /rr /sa

- ✓ d'abord la variable de maladie ou du critère de jugement : ici l'arrêt de l'allaitement à 1 mois
- ✓ /r pour avoir les % en lignes (+++ nécessaire pour le risque relatif)
- ✓ / t pour le test du  $\chi^2$
- ✓ / rr pour le risque relatif
- ✓ / sa pour avoir le codage avec 1 avant 2

Outcome:arretm1						
Moment	1 (Arret/1 mois)	%	2 (Pas arret/1 mois)	%	Total	%
1 (Pendant la G)	11	(19.6)	45	(80.4)	56	(100.0)
2 (Avant la G)	10	(6.0)	157	(94.0)	167	(100.0)
Total	21	(9.4)	202	(90.6)	223	
Percents: (Row) Exposure: Moment = 0						
Outcome: arretm1 = 1 Chi2= 9.167 df(1) p= 0.0025						
<b>RR = 3.28 (95% CI: 1.47-7.31)</b>						

Le RR est un rapport d'incidence : il se calcule de la façon suivante à partir du tableau 2 \* 2 :

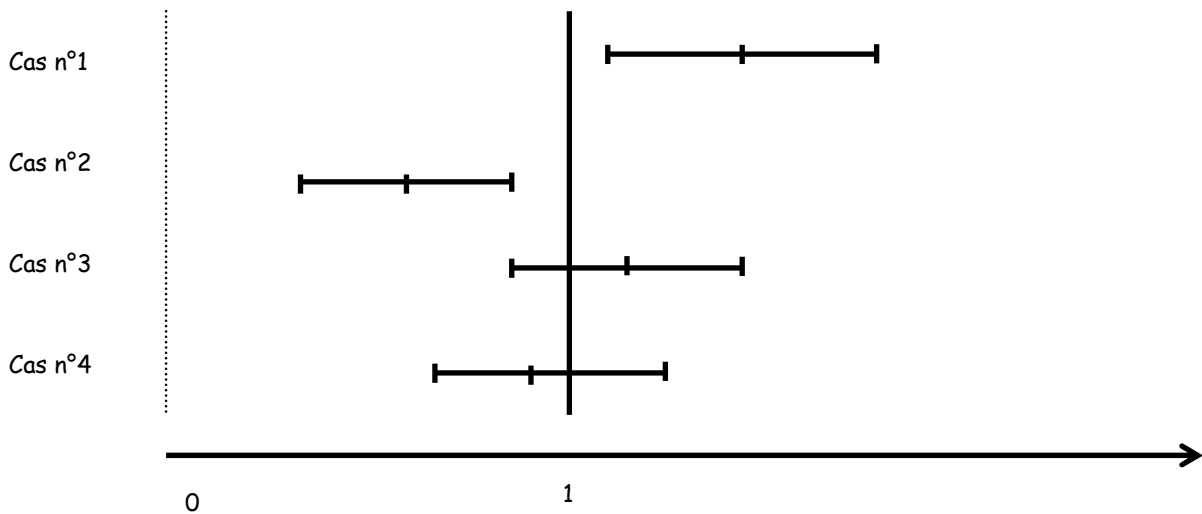
- incidence de arret1 lorsque la décision d'allaitement a été pris PENDANT la grossesse : 11 / 56 = 19.6 %
  - incidence de arret1 lorsque la décision d'allaitement a été pris AVANT la grossesse : 10 / 167 = 6.0 %
- soit RR = 19.6 / 6.0 = 3.28

Le RR peut être assorti d'un intervalle de confiance à 95 % : s'il comprend 1, il n'est pas significatif ; s'il ne comprend pas 1, il est significatif à  $p < 0.05$ . Ici, l'IC à 95 % vaut 1.47 - 7.31 et est différent de 1 (au sens statistique).

A signaler plusieurs points :

- RR se lit en ligne
- Les codages de la maladie (outcome) et de l'exposition doivent être binaires. En cas de codage en 3 classes par exemple, faire deux RR en 2 tableaux avec des « select ».
- Association ne veut pas dire « cause ».

Figure 6 bis : Valeur et intervalle de confiance d'un RR : exemple fictif de 4 cas différents



Pour le cas n°1, le RR > 1 et significatif (c'est à dire significativement différent de 1). Pour le cas n°2, le RR < 1 et significatif. Pour le cas n°3, le RR > 1 et non significatif. Pour le cas n°4, le RR < 1, et non significatif.

Autre exemple avec RR = 0.41 , ici NS car comprend 1 :

Outcome:arretm1						
COUPLE	1	%	2	%	Total	%
3 (cadre)	5	(5.2)	92	(94.8)	97	(100.0)
8 (pas cadre)	16	(12.7)	110	(87.3)	126	(100.0)
Total	21	(9.4)	202	(90.6)	223	

Percents: (Row) Exposure: COUPLE = 3  
 Outcome: arretm1 = 1 Chi2= 3.656 df(1) p= 0.0559  
 RR = 0.41 (95% CI: 0.15-1.07)

**Les tableaux dans les enquêtes d'incidence** : la mesure est la proportion de « malades » (critère de jugement) +++ pour les variables qualitatives et les valeurs pour les malades et les non-malades pour les variables quantitatives.

**Tableau I : Présentation d'un tableau dans les enquêtes d'incidence**

Variables qualitatives	Incidence de l'arrêt d'allaitement à 1 mois (%)	RR (IC à 95 %)	p
Age < 30 ans	17.6 %	4.64	< 10 <sup>-4</sup>
≥ 30 ans	3.8 %	(1.76 - 12.22)	
Primipares	14.1 %	2.31	0.043
Multipares	6.1 %	(1.00 - 5.36)	
Mère allaitée : oui	5.6 %	0.47	0.11
non	11.9 %	(0.18 - 1.24)	
IMC > 30	15.4 %	1.70	0.44
IMC ≥ 30	9.0 %	(0.44 - 6.53)	
Reprise du travail : oui	5.6 %	0.35	0.01
non	16.0 %	(0.15 - 0.81)	

Variables quantitatives	Arrêt allaitement à 1 mois	Pas d'arrêt d'allaitement à 1 mois	p
Age (ans)	27.2 ± 4.4	31.2 ± 4.7	0.004
IMC (kg/m <sup>2</sup> )	23.9 ± 4.1	22.3 ± 3.9	0.06

*La ligne suivante n'est pas l'objectif d'une enquête de cohorte*

*Car c'est le taux d'exposition selon la « maladie » ou non, ici taux de mères jeunes en cas d'arrêt versus en l'absence d'arrêt.*

Age < 30 ans (%)	76.2	37.7	< 0.001
------------------	------	------	---------

### C. Les enquêtes cas-témoins → OR

Dans ce type d'enquêtes, on choisit des cas MALADES ou CAS. Exemple : âge des mères (en années) avec la commande (par exemple des infections nosocomiales) et on y associe des NON MALADES ou TEMOINS ou CONTROLS (anglicisme) (1 témoin pour 1 cas par exemple). Ce peut être enquête faite ad hoc, ou au sein d'une autre enquête. Cette catégorie d'enquête est plutôt utilisée quand les cas sont rares et que l'on veut rechercher des facteurs de risques (ou facteurs d'exposition).

Exemple : Macrosomes= cas (4000g et plus à terme) comparés à des eutrophes = témoins (3500 - 3999 g à terme). 79 cas versus 71 témoins.

→ Exemples : tables castem episio /c /t /o

castem						
Episio	1 = Cas	%	2 = Témoins	%	Total	%
1 (oui)	22	{37.9}	20	{35.7}	42	{36.8}
2 (non)	36	{62.1}	36	{64.3}	72	{63.2}
Total	58	{100.0}	56	{100.0}	114	
Percents: {Col} Exposure: Episio = 1						
Outcome: castem = 1 Chi2= 0.060 df(1) p= 0.8062						
<b>Odds Ratio = 1.10 (95% CI: 0.48-2.53)</b>						

Il n'y a pas de différence entre les cas et les témoins pour le taux d'épisiotomies (37.9 % versus 35.7 %) avec un p = 0.80 et un OR = 1.10 non significatif.

On essaie de voir si le périmètre abdominal est associé à la macrosomie. Hypothèse de seuil à 350 mm

→ Exemples

```
define pa350 #
if pa > 350 then pas350=1 else pa350=2
tables castem pa360 /c /t /o
```

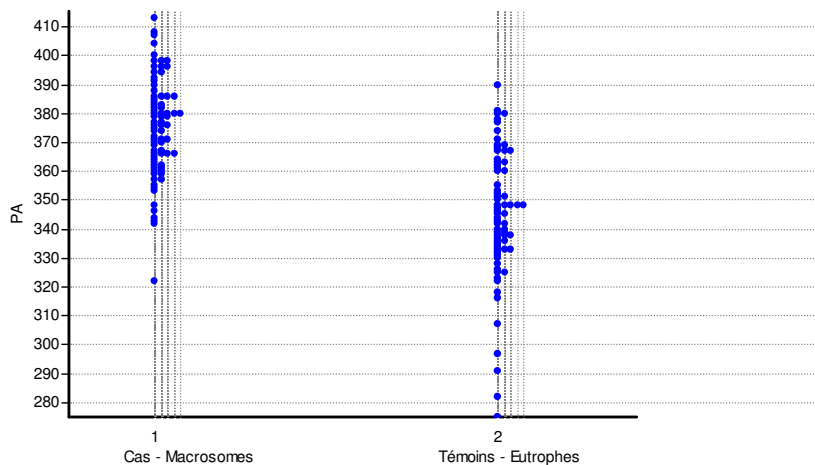
castem						
pa350	1	%	2	%	Total	%
1 (> 350 mm)	64	{91.4}	26	{40.6}	90	{67.2}
2 (≤ 350 mm)	6	{8.6}	38	{59.4}	44	{32.8}
Total	70	{100.0}	64	{100.0}	134	

Percents: {Col} Exposure: pa350 = 1  
 Outcome: castem = 1 Chi2= 39.127 df(1) p= 0.0000  
**Odds Ratio = 15.22 (95% CI: 5.53-49.47)**

La PA au-dessus de 350 est associé significativement à la macrosomie (OR = 15.2) avec  $p < 10^{-4}$ .  
 Remarque : la moyenne de PA pour les macrosomes est de  $373 \pm 17$  mm versus  $347 \pm 21$  mm ( $p < 10^{-6}$ ).

Exemples : dotplot castem PA /edit

Figure 7 : Dot plot de valeurs en vue de la décision



Les commentaires :

- = Le sens du tableau n'a pas importance contrairement au tableau pour le RR
  - = Un OR ne peut se calculer que sur un tableau 2 x 2 : nécessité d'établir des regroupements ou des "cut points". L'OR est le même si les lignes et les colonnes sont inversées (au contraire du RR).
  - = Association ne veut pas dire cause ++++
- Les tableaux dans les enquêtes cas-témoins : la mesure est le taux d'exposition chez les cas et chez les témoins +++. Deux parties : le premier tableau montre que les cas et les témoins sont identiques (pas de p significatif) : pas de OR et souvent pas de « p » montré.

Tableau II : Tableau en cas d'enquêtes cas-témoins (comparaison préalable)

Les deux populations doivent être identiques	Cas N=79	Témoins N=71	p
Age mère (ans)	30.2 ± 5.4	29.3 ± 5.1	0.31
Primiparité (%)	31.4	45.3	0.09
Sexe (% garçons)	62.9	50.0	0.13

Le deuxième tableau montre les différences recherchées avec OR (pour les variables quantitatives) et « p ».

Tableau III : Tableau en cas d'enquêtes cas-témoins (taux d'exposition ou moyenne selon les cas/témoins)

Les critères de jugement	Cas n=79	Témoins n=71	OR IC à 95 %	P
Périmètre abdominal (mm)	374 ± 17	344 ± 23		< 10 <sup>-6</sup>
PA > 350 mm (%)	91.4	40.6	15.22 5.53 - 49.47	< 10 <sup>-4</sup>
Césarienne (%)	26.6	21.1	1.5 0.59 - 3.12	0.43
Épisio (%) pour les voies basses	N=58 37.9	N=56 35.7	1.10 0.48 - 2.53	0.80

#### D. Les études de dépistage ou de diagnostic

Voir OPEN-EPI pour l'étude des Se, Sp, VPP, VPN... Courbes ROC : un peu OPEN-EPI, MEDCALC ou SPSS.

#### E. Les courbes de survie

Une étude dite de « survie » correspond à des situations où on dispose de données dans le temps (en jours, semaines, mois, années), et une donnée dite « événement » qui est classiquement le décès, mais qui peut-être l'entrée dans la maladie, l'arrêt de l'allaitement, l'accouchement.... Les données de temps peuvent être incomplètes (manque de recul, perdus de vue) : elles sont dites censurées ; elles comptent alors pour le temps où elles sont connues.

On part de 100 % de personnes « en vie » : à chaque décès, la probabilité d'être en vie baisse, à chaque perdu de vue, la probabilité reste la même, mais le dénominateur change. Les probabilités de survie sont dites proportionnelles dans le sens où les probabilités de survie se multiplient au fil du temps.

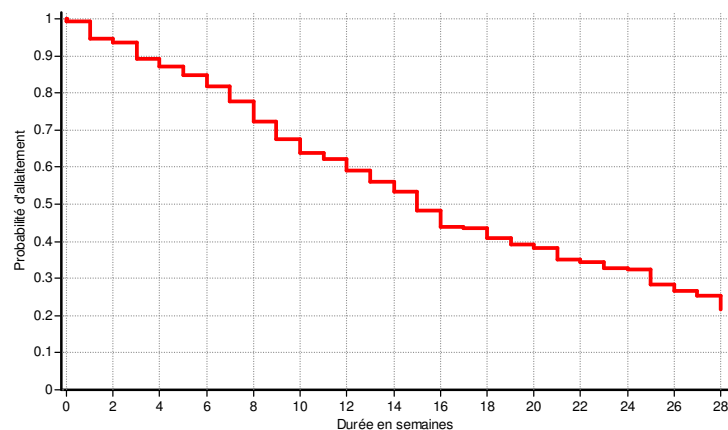
EPIDATA fait état des résultats suivants :

- ✓ Table de survie avec les probabilités à chaque temps
- ✓ Médiane de survie : temps au bout duquel 50 % des sujets sont en survie
- ✓ Courbe de survie qui part de 100 % et qui est en escalier à chaque chute de probabilité
- ✓ En cas de comparaison de courbes : test du logrank qui teste la différence entre les deux courbes sous condition qu'elles ne se croisent pas, « hazard ratio » ou risque relatif de « décès ».

Les exemples sont tirés d'une enquête sur la durée d'allaitement.

➔ Exemple : lifetable arret durall /ncoci /p50 /edit

Figure 8 : Courbe de survie



Life Table: ARRET					
Total N	Cases n	Time Minimum	Maximum	Sum time	Time S=0.50
239	162	0	42	3500	<b>15.00</b>
Outcome: ARRET = 1					

La médiane de survie est de 15 semaines.

Interval	N At risk	Deaths	Lost	Survival
0-1	239	2	0	0.99
1-2	237	11	0	0.95
2-3	226	2	0	0.94

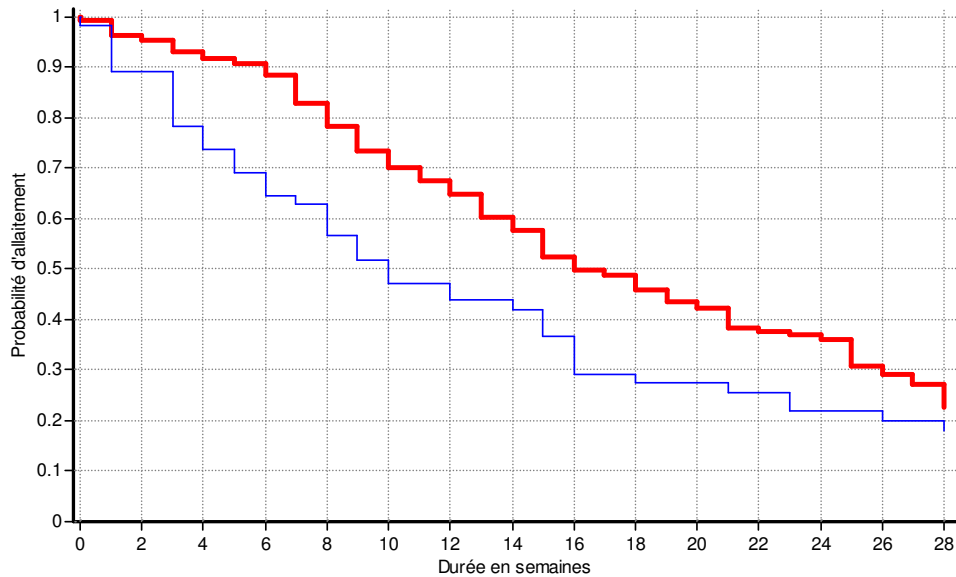
12-13	132	6	0	0.59
13-14	126	7	6	0.56
14-15	113	5	5	0.53
15-16	103	10	1	0.48
16-17	92	8	0	0.44
17-18	84	1	1	0.44

26-27	50	3	7	0.27
27-28	40	2	11	0.25
28-29	27	27	4	0.22

Warning: Time variable has type float, interval=1 assumed  
Time intervals from left number up to right number

➔ Exemple : lifetable arret durall /ncoci /p50 /edit /by=moment

Figure 9 : Courbes de survie selon un critère à deux classes



Life Table: ARRET by Moment

	Total N	Cases n	Expected Cases	Time Minimum	Maximum	Sum time	Hazard Ratio	Time S=0.50	Time Diff.
1	174	112	125.09	0	42	2702	Ref.	16.00	Ref.
2	65	50	36.91	0	34	798	1.62	10.00	-6.00

Outcome: ARRET = 1  
 Log Rank test of equality of survivor function: Chi2(1)= 6.017 P= 0.0142  
 LR test of homogeneity among groups Chi2(1)= 5.570 p= 0.0183

Les deux courbes sont significativement différentes avec p=0.01, et le risqué relative est de 1.62 (en cas de moment code 2, il y a 1.62 fois plus d'arrêt d'alalitement.

Interval	NA+ risk	Deaths	Lost	Survival
<b>Moment = 1</b>				
0-1	174	1	0	0.99
1-2	173	5	0	0.97
2-3	168	2	0	0.95
3-4	166	4	0	0.93

-----

25-26	47	7	2	0.31
26-27	38	2	6	0.29
27-28	30	2	10	0.27
28-29	18	3	15	0.23

Interval	NA+ risk	Deaths	Lost	Survival
<b>Moment = 2</b>				
0-1	65	1	0	0.98
1-2	64	6	0	0.89

=====

21-22	15	1	0	0.26
23-24	14	2	0	0.22
26-27	12	1	1	0.20
27-28	10	0	1	0.20
28-29	9	1	8	0.18

*Warning: Time variable has type float, interval=1 assumed  
 Time intervals from left number up to right number*

## V. Pour conclure

### *A. Ouvrages de références*

#### \* Facile

- Epidémiologie sans peine. Goldberg
- Le Jeu de la Science et du hasard : la statistique et le vivant. D. Schwartz
- Décision médicale. Bernard Grenier. Masson (réédition 1996)
- La recherche clinique (de l'idée à la publication). Gilles Landrivon. Masson
- **Statistique épidémiologique. T. Ancelle. Maloine +++**

#### \* Classique

- Statistique appliquée à la médecine. D. Schwartz. Flammarion Méd.-Sc.

#### \* Plus difficile

- Recherche Clinique. A. Laplanche, C. Coum-Nogué. Flammarion.
- Epidémiologie : méthodes et pratiques. C Rumeau-Rouquette. Flammarion
- Méthodes statistiques en médecine et biologie. Jean Bouyer. Ed INSERM 1996
- Les statistiques dans les sciences de la vie. Bruno Falissard. Masson 1996

### *B. La partie statistique d'une étude n'est pas la plus importante*

#### **Avant de commencer :**

- Quel objectif ?
- Que veut-on mesurer ? Comment le mesurer ? Quelle forme de variables ?
- Combien de sujets sont nécessaires ?

#### **A l'analyse :**

- Bien comprendre l'analyse univariée
- Bien analyser les liens entre les variables d'exposition : analyse des biais.

## VI. Tests statistiques usuels

Voir aussi (plus complet) : C:\Program Files\OpenEpi\Menu\OpenEpiMenu.htm → Choisir une méthode

En gras les commandes de EPIDATA 2.x

vqual = variables qualitatives et vquant = variables quantitatives

Type de variables	Description	Comparaison	
		<i>Tests paramétriques (nombre suffisant ou répartition normale)</i>	<i>Tests non paramétriques (nombre insuffisant ou répartition non normale)</i>
<b>Une proportion</b>	Pourcentage avec n et (%) ou intervalle de confiance à 95 % <b>freq vqual /c /ci</b>	/ proportion théorique Test de $\epsilon$ (à la main)	Loi de Poisson (à la main)
<b>Deux proportions</b>	Différence de proportion avec IC à 95 %	Test du $\chi^2$ avec 2 ddl (df) <b>tables vqual1 vqual2 /r /t</b>	Test de Fisher <b>tables vqual1 vqual2 /r /ex</b>
<b>Plus de deux proportions</b>	---	Test de $\chi^2$ avec x ddl (df) <b>tables vqual1 vqual2 /r /t</b>	---
<b>Une moyenne</b>	Moyenne $\pm$ un écart-type ou IC95 % (- 2 ET + 2 ET) <b>means vquant</b>	/ moyenne théorique : test de $\epsilon$ (à la main) / 0 : <b>means vquant /t</b>	
<b>Deux moyennes</b>	Différence de moyenne avec IC à 95 % <b>means diff /t</b>	Test de t de Sudent avec N-2 ddl (df) <b>means vquant vqual /t</b>	Test de Wilcoxon ou Mann-Whitney <b>kwallis vquant vqual</b>
<b>Plus de deux moyennes</b>		ANOVA <b>means vquant vqual /t</b>	Test de Kruskal-Wallis <b>kwallis vquant vqual</b>
<b>Deux quantités</b>		Droite de régression avec le test de la pente / pente = 0 <b>regress vquant1 vquant2</b> Coefficient de corrélation r de Pearson <b>correlate vquant1 vquant2</b>	Coefficient de corrélation $\rho$ de Spearman
<b>Survie</b>	Courbe (probabilité de survie) de Kaplan-Meier ou actuarielle avec médiane de survie <b>lifetable évènement durée /p50</b>	Comparaisons de courbes Test du logrank <b>lifetable évènement durée /p50 /t</b>	Comparaisons de courbes Test du logrank <b>lifetable évènement durée /p50 /t</b>
<b>Concordance</b> - 2 observateurs - Plus de 2 observateurs		Test de kappa de Cohen (voir Open Epi)	

## VII. Tests et conditions de validité (pages du livre de D. SCHWARTZ)

### I- DONNEES QUALITATIVES

#### 1. Intervalle de confiance d'un pourcentage observé $p_o$

- nps, nqs, npi, nqi  $\geq 5$  (p 47)
- sinon utiliser la table 2A (p 290)

#### 2. Comparaison d'un % théorique $p$ à un % observé (méthode du $\chi^2$ )

- np, nq  $\geq 5$ , ou encore si les effectifs calculés sont supérieurs à 5 (p 41)
- sinon, par la méthode du  $\chi^2$ , correction de la différence avec  $-\frac{1}{2}$  (p 70)

#### 3. Comparaison de deux pourcentages indépendants par le test du $\chi^2$

- les effectifs calculés  $\geq 5$ ,  $\chi^2$  habituel (p 84)
- les effectifs calculés  $< 5$  et  $> 3$ ,  $\chi^2$  avec la correction de Yates avec  $-\frac{1}{2}$  (p 94)
- les effectifs calculés  $\leq 3$ , test de Fisher (p 97)

#### 4. Comparaison de deux pourcentages de séries appariées par le test du $\chi^2$

- le nombre de paires discordantes  $\geq 10$ , utiliser  $\chi^2$  apparié (p 64)
- sinon correction de -1 de la différence du  $\chi^2$  apparié (p 95)

#### 5. Comparaison de plusieurs pourcentages observés

- les effectifs calculés  $\geq 5$  (p 81)
- sinon regrouper les effectifs (sous conditions que cela corresponde à l'objectif visé, et/ou à une validité biologique)

### II- DONNEES QUANTITATIVES

#### 6. Intervalle de confiance d'une moyenne d'un individu

- la variable suit une loi normale, et  $n \geq 30$ , utiliser  $\epsilon$  (devoir n° 7, question 4) (p135)
- la variable suit une loi normale, et  $n < 30$ , utiliser t avec (n-1) ddl (p 154)

#### 7. Intervalle de confiance d'une moyenne observée

- $n \geq 30$ , utiliser  $\epsilon$  (p 133)
- $n < 30$ , et la variable suit une loi normale, utiliser t avec (n-1) ddl (p 154)

#### 8. Comparaison d'une moyenne observée à une moyenne théorique

- $n \geq 30$ , utiliser  $\epsilon$  (p 139)
- $n < 30$ , et la variable suit une loi normale, utiliser t avec (n-1) ddl (p 156)

#### 9. Comparaison de deux moyennes

- $n_A$  et  $n_B \geq 30$ , utiliser  $\epsilon$  (p 143)
- $n_A$  ou  $n_B < 30$ , et populations de distribution normales et de même variance, utiliser t avec  $(n_A + n_B - 2)$  ddl (p 158)
- $n < 30$ , et la variable suit une loi qcq, utiliser un test non-paramétrique

#### 10. Comparaison de deux moyennes de séries appariées (n couples)

- $n \geq 30$ , utiliser  $\epsilon$  (p 150)
- $n < 30$ , et la différence suit une loi normale, utiliser t avec (n-1) ddl (p 161)

#### 11. Comparaison de 2 variances de séries indépendantes

- les deux séries sont extraites de populations à distribution normale (p 168)

#### 12. ANOVA

- distributions normales dans chaque échantillon et de même variance (p 164)

#### 13. Coefficient de corrélation, pente de la droite de régression

- l'une au moins des distributions est normale, et de variance constante (p 218)