

# Projet CORUS - Obe-Maghreb

---

## Ecole thématique « Gestion et analyse de données d'enquêtes épidémiologiques »

20-29 avril 2010, Rabat, Maroc

## Analyse de données Exercices pratiques avec EpiData Analysis

Pierre Traissac

IRD

UMR 204 NUTRIPASS IRD, UM1, UM2

« Prévention des malnutritions et pathologies associées »

Montpellier, France



## **Analyse de données d'enquêtes épidémiologiques Exercices pratiques avec EpiData Analysis**

### **4. Dictionnaire des variables (incluant les variables dérivées)**

N.B. : Les fichiers de données des exercices sont un sous-ensemble de la base de donnée de l'enquête transversale nationale de nutrition Tunisie 1996/97, propriété de l'INNTA (Institut National de Nutrition de Tunisie, Tunis, Tunisie). Ces données sont la propriété de l'INNTA et mises à disposition des utilisateurs dans le cadre des exercices proposés, à l'exclusion de toute autre utilisation.

### Table « ménages » : unité statistique ménage

- fichier format Epidata : **menages2010\_2ana.rec** ( n=1763 observations, p=44 variables)

- date de l'état : 10/04/2010

Numéro d'ordre dans le fichier	Nom de la variable	Format	Contenu	Unités ou codes	n	Minimum - maximum	Type statistique	Remarques diverses
1	<b>idmen</b>	Numérique	Clé primaire : identifiant unique de ménage	sans objet	1763	110701 à 832406	identifiant	Combine n° de gouvernorat, de délégation et de ménage
2	<b>gouvern</b>	Numérique	Gouvernorat	Détails non donnés dans le cadre de l'exercice	1763	11 à 83	qualitatif	
3	<b>deleg</b>	Numérique	Numéro de délégation dans le gouvernorat	Détails non donnés dans le cadre de l'exercice	1763	1 à 72	qualitatif	
4	<b>numen</b>	Numérique	Numéro de ménage dans la délégation	Numéro séquentiel	1763			
5	<b>datenq</b>	Date	Date de l'enquête		1763	13 jan 1996 au 28 dec 1997		Calculée à partir des variables initiales <i>datevisj</i> , <i>datevism</i> , <i>datevisa</i>
6	<b>stratech</b>	Numérique	Strates de l'échantillonnage	Détails non donnés dans le cadre de l'exercice	1763	111 à 832	qualitatif	Combine <i>gouvern</i> et <i>milieu</i>
7	<b>gouvdel</b>	Numérique	Grappes de l'échantillonnage	Détails non donnés dans le cadre de l'exercice	1763	1107 à 8324	qualitatif	Combine <i>gouvern</i> et <i>deleg</i>
8	<b>pond1</b>	Numérique	Poids de sondage lié à la stratification			316 à 2690		Car probabilités de tirage au sort ménages inégales. Est constante par <i>stratech</i> .
9	<b>region</b>	Numérique	Région administrative	1 : Grand Tunis 2 :Nord Est 3 :Nord Ouest 4 :Centre Ouest 5 : Centre Est 6 : Sud Ouest 7 : Sud Est	1763	1 à 7	qualitatif	Est un recodage de <i>gouvern</i>
10	<b>habitat</b>	Numérique	Type d'habitat	1 : grandes villes 2 : autres communes 3 : rural aggloméré 4 :rural dispersé	1763	1 à 4	qualitatif	Selon classification en vigueur lors de l'enquête
11	<b>milieu</b>	Numérique	Type d'habitat	1 : urbain 2 : rural	1763	1 à 2	qualitatif	Est un recodage de <i>habitat</i>
12	<b>typeloge</b>	Numérique	Type de logement	1 :villa 2 :appartement 3 :studio 4 :maison arabe 5 :gourbi 6 :autre	1703	1 à 6	qualitatif	
13	<b>mode</b>	Numérique	Mode d'occupation du logement	1 : propriétaire 2 :locataire 3 : logé à titre gratuit 4 : accédant à la propriété	1702	1 à 4	qualitatif	
14	<b>nbchamb</b>	Numérique	Nombre de chambres du logement	Nombre de chambres	1529	1 à 9	quantitatif	
15	<b>bain</b>	Numérique	Salle de bain dans logement	1 :oui 2 :non	1743	1 à 2	dichotomique	
16	<b>douche</b>	Numérique	Douche dans logement	1 :oui 2 :non	1744	1 à 2	dichotomique	
17	<b>salleau</b>	Numérique	Salle d'eau dans logement	1 :oui 2 :non	1742	1 à 2	dichotomique	
18	<b>toilette</b>	Numérique	Toilettes dans logement	1 :oui 2 :non	1735	1 à 2	dichotomique	
19	<b>robinet1</b>	Numérique	Eau par robinet dans logement	1 :oui 2 :non	1739	1 à 2	dichotomique	
20	<b>puits</b>	Numérique	Eau à partir d'un puits	1 :oui 2 :non	1739	1 à 2	dichotomique	
21	<b>robinet2</b>	Numérique	Eau par robinet public	1 :oui 2 :non	1735	1 à 2	dichotomique	
22	<b>égout</b>	Numérique	Evacuation eaux usées par égout	1 :oui 2 :non	1739	1 à 2	dichotomique	
23	<b>fosse</b>	Numérique	Evacuation eaux usées par fosse sceptique	1 :oui 2 :non	1727	1 à 2	dichotomique	

Numéro d'ordre dans le fichier	Nom de la variable	Format	Contenu	Unités ou codes	N	Minimum - maximum	Type statistique	Remarques diverses
24	<b>electric</b>	Numérique	Logement raccordé au réseau électrique	1 :oui 2 :non	1744	1 à 2	dichotomique	
25	<b>sonede</b>	Numérique	Logement raccordé réseau sonede (eau)	1 :oui 2 :non	1736	1 à 2	dichotomique	
26	<b>frigo</b>	Numérique	Logement équipé d'un réfrigérateur	1 :oui 2 :non	1723	1 à 2	dichotomique	
27	<b>phone</b>	Numérique	Logement équipé d'un téléphone	1 :oui 2 :non	1738	1 à 2	dichotomique	
28	<b>tele</b>	Numérique	Logement équipé d'une télévision	1 :oui 2 :non	1738	1 à 2	dichotomique	
29	<b>video</b>	Numérique	Logement équipé d'un magnétoscope	1 :oui 2 :non	1743	1 à 2	dichotomique	
30	<b>cuisine</b>	Numérique	Logement équipé d'une cuisinière	1 :oui 2 :non	1741	1 à 2	dichotomique	
31	<b>lavess</b>	Numérique	Logement équipé d'un lave vaisselle	1 :oui 2 :non	1750	1 à 2	dichotomique	
32	<b>clime</b>	Numérique	Logement équipé d'un climatiseur	1 :oui 2 :non	1745	1 à 2	dichotomique	
33	<b>central</b>	Numérique	Logement équipé du chauffage central	1 :oui 2 :non	1751	1 à 2	dichotomique	
34	<b>parabole</b>	Numérique	Logement équipé d'une parabole	1 :oui 2 :non	1750	1 à 2	dichotomique	
35	<b>lavlinge</b>	Numérique	Logement équipé d'un lave linge	1 :oui 2 :non	1745	1 à 2	dichotomique	
36	<b>menage</b>	Numérique	Présence d'une aide ménagère	1 :oui 2 :non	1751	1 à 2	dichotomique	
37	<b>voiture</b>	Numérique	Ménage possède une voiture	1 :oui 2 :non	1742	1 à 2	dichotomique	
38	<b>confort</b>	Numérique	Score de confort matériel du ménage	Nombre d'éléments de confort matériel	1671	0 à 10	quantitatif	A partir des 12 variables de de possessions de frigo à voiture
39	<b>confc3</b>	Numérique	Classes de score de confort du ménage	1: niveau de confort faible(0 à2) 2: niveau de confort moyen (3) 3: niveau de confort élevé (4à10)	1671	1 à 3	qualitatif	Recodage de confort
40	<b>eco1</b>	Numérique	Indice de niveau économique du ménage	Sans unité (classement des ménages par indice de niveau de vie croissant)	1725	0 à 100	quantitatif	Est la coordonnée du ménage sur le premier axe de l'AFC sur un ensemble de variables d'équipement et de niveau de vie du ménage. Utilise typologie, mode et les 23 variables de bain à voiture
41	<b>eco1c3</b>	Numérique	Terciles de niveau économique	1 : 1 <sup>er</sup> tercile « bas » 2 : 2 <sup>ème</sup> tercile « moyen » 3 : 3 <sup>ème</sup> tercile « élevé »	1725	1 à 3	qualitatif	Recodage de eco1. Les terciles sont estimés sur l'ensemble des ménages.
42	<b>eco1c3b</b>	Numérique	Terciles de niveau économique par milieu (rural vs urbain)	1 : 1 <sup>er</sup> tercile « bas » 2 : 2 <sup>ème</sup> tercile « moyen » 3 : 3 <sup>ème</sup> tercile « élevé »	1725	1 à 3	qualitatif	Recodage de eco1. Les terciles sont estimés séparément dans chaque milieu (rural vs urbain)
43	<b>depenses</b>	Numérique	Dépenses alimentaires mensuelles	Dinars courants à la date d'enquête	1508	20 à 650	quantitatif	Déclaratif
44	<b>depc3</b>	Numérique	Classes de dépenses alimentaires mensuelles	1 : <100 dinars 2 : 100 à <200 dinars 3 : >=200 dinars	1508	1 à 3	qualitatif	Recodage de depenses

**Table « menfem » : unité statistique femme de 20 à 59 ans , non enceinte, non allaitante ( combine variables femme et ménage)**

- fichier format Epidata : **menfem2010\_2ana.rec**, ( n=1849 observations, p=77 variables)

- date de l'état : 10/04/2010

Numéro d'ordre dans le fichier	Nom de la variable	Format	Contenu	Unités ou codes	n	Minimum - maximum	Type statistique	Remarques diverses
1	<b>idfem</b>	Numérique	Clé primaire : identifiant unique de femme	sans objet	1849	1107011 à 8324061	identifiant	Combine n° de gouvernorat, de délégation, de ménage et de femme
2	<b>idmen</b>	Numérique	Clé secondaire : identifiant de ménage	sans objet	1849	110701 à 832406	identifiant	Combine n° de gouvernorat, de délégation et de ménage
3	<b>gouvern</b>	Numérique	Gouvernorat	Détails non donnés dans le cadre de l'exercice	1849	11 à 83	qualitatif	
4	<b>deleg</b>	Numérique	Numéro de délégation dans le gouvernorat	Détails non donnés dans le cadre de l'exercice	1849	1 à 72	qualitatif	
5	<b>numen</b>	Numérique	Numéro de ménage dans la délégation	Numéro séquentiel	1849	1 à 22	qualitatif	
6	<b>nufem</b>	Numérique	Numéro de femme dans le ménage	Numéro séquentiel	1849	1 à 5	qualitatif	
7	<b>datenq</b>	Date	Date de l'enquête		1849	13 jan 1996 au 24 dec 1997		Calculée à partir des variables initiales <i>datevisj</i> , <i>datevism</i> , <i>datevisa</i>
8	<b>stratech</b>	Numérique	Strates de l'échantillonnage	Détails non donnés dans le cadre de l'exercice	1849	111 à 832	qualitatif	Combine <i>gouvern</i> et <i>milieu</i>
9	<b>gouvdel</b>	Numérique	Grappes de l'échantillonnage	Détails non donnés dans le cadre de l'exercice	1849	1107 à 8324	qualitatif	Combine <i>gouvern</i> et <i>deleg</i>
10	<b>pond1</b>	Numérique	Poids de sondage lié à la stratification	sans objet	1849	316 à 2690		Car probabilités de tirage au sort ménages inégales. Est constante par <i>stratech</i> et ménage.
11	<b>pond2</b>	Numérique	Poids pour redressement non réponses	sans objet	1849	0.56 à 1.77		Postratification : redressement de la structure d'âge et sexe par région
12	<b>region</b>	Numérique	Région administrative	1 : Grand Tunis 2 :Nord Est 3 :Nord Ouest 4 :Centre Ouest 5 : Centre Est 6 : Sud Ouest 7 : Sud Est	1849	1 à 7	qualitatif	Est un recodage de gouvern
13	<b>habitat</b>	Numérique	Type d'habitat	1 : grandes villes 2 : autres communes 3 : rural aggloméré 4 :rural dispersé	1849	1 à 4	qualitatif	Selon classification en vigueur lors de l'enquête
14	<b>milieu</b>	Numérique	Type d'habitat	1 : urbain 2 : rural	1849	1 à 2	qualitatif	Est un recodage de habitat
15	<b>jnais</b>	Numérique	Jour de naissance	sans objet	1849	1 à 31		
16	<b>mnais</b>	Numérique	Mois de naissance	sans objet	1849	1 à 12		
17	<b>anais</b>	Numérique	Année de naissance	sans objet	1849	1936 à 1977		
18	<b>datenais</b>	Date	Date de naissance		1849	25 juil 1936 au 23 oct 1977		Calculé à partir de <i>jnais</i> , <i>mnais</i> , <i>anais</i>
19	<b>age</b>	Numérique	Age	Années décimales	1849	20,02 à 59.98	quantitatif	Calculé à partir de <i>datenq</i> et <i>datnai</i>
20	<b>agec4</b>	Numérique	Age en classes	1 :20<=age<30 2 :30<=age<40 3 :40<=age<50 4 :50<=age<60	1849	1 à 4	qualitatif	Recodage de age
21	<b>matc3</b>	Numérique	Statut matrimonial	1 :célibataire 2 :mariée 3 :veuve ou divorcée	1849	1 à 3	qualitatif	Est un recodage de la variable initiale <i>etatcivi</i> à 4 modalités par regroupement de veuves et divorcées
22	<b>scoc4</b>	Numérique	Niveau scolaire	1 :analphabète 2 :primaire 3 : secondaire 4 :supérieur	1814	1 à 4	qualitatif	Est un recodage de la variable initiale <i>niveau</i> à 6 modalités
23	<b>scoc3</b>	Numérique	Niveau scolaire	1 :analphabète 2 :primaire 3 : secondaire ou supérieur	1814	1 à 3	qualitatif	Recodage de <i>scoc4</i>
24	<b>actpc2</b>	Numérique	Activité professionnelle	1 :oui 2 :non	1842	1 à 2	dichotomique	Est un recodage de la variable initiale <i>profess</i> contenant des codes profession
25	<b>parc4</b>	Numérique	Parité	1 :0 2 :1,2 3 :3,4 4 :5 et +	1813	1 à 4	qualitatif	Est un recodage de la variable initiale <i>nbgross</i> (nombre de grossesses)

Numéro d'ordre dans le fichier	Nom de la variable	Format	Contenu	Unités ou codes	n	Minimum - maximum	Type statistique	Remarques diverses
26	<b>poids</b>	Numérique	Poids	kg	1837	32.7 à 159	quantitatif	
27	<b>taille</b>	Numérique	Taille	cm	1836	113 à 182	quantitatif	
28	<b>imc</b>	Numérique	Indice de masse corporelle	kg/m <sup>2</sup>	1835	13.70 à 53.13	quantitatif	$Imc=poids/(taille/100)^2$
29	<b>imc_c4</b>	Numérique	Classes de corpulence	1 : « maigre » $imc < 18.5$ 2 : « normal » $18.5 \leq imc < 25$ 3 : « surpoids » $25 \leq imc < 30$ 4 : « obésité » $30 \leq imc$	1835	1 à 4	Qualitatif	Recodage de imc
30	<b>obesite</b>	Numérique	Obésité (oui/non)	1 : $imc \geq 30$ 0 : $imc < 30$	1835	0 à 1	Dichotomique	Recodage de imc
31	<b>tt</b>	Numérique	Tour de taille	cm	1738	57 à 135	quantitatif	
32	<b>th</b>	Numérique	Tour de hanche	cm	1738	63 à 154	quantitatif	
33	<b>rth</b>	Numérique	Rapport tour de taille/tour de hanche	Sans unité	1738	0.65 à 1.38	quantitatif	$th=tt/th$
34	<b>tas</b>	Numérique	Tension artérielle systolique	mm hg	1825	80 à 240	quantitatif	
35	<b>tad</b>	Numérique	Tension artérielle diastolique	mm hg	1823	30 à 140	quantitatif	
36	<b>hemoglob</b>	Numérique	Hémoglobine	g/dl	1789	5.3 à 22.2	quantitatif	
37	<b>glycemie</b>	Numérique	Glycémie à jeun	mmol/l	1771	2.5 à 51	quantitatif	
38	<b>cholest</b>	Numérique	Cholestérolémie totale	mmol/l	1657	0.61 à 9.96	quantitatif	
39	<b>trigly</b>	Numérique	Triglycéridémie	mmol/l	1652	0.23 à 8.18	quantitatif	
40	<b>cal</b>	Numérique	Apport énergétique total / jour	kcal	1819	818 à 3832.73	quantitatif	
41	<b>calp</b>	Numérique	Apport énergétique total / jour rapporté aux besoins	sans unité	1818	0.36 à 1.76	quantitatif	
42	<b>glucide1</b>	Numérique	Apport en glucides en % de l'apport total	sans unité	1819	43.2 à 71	quantitatif	
43	<b>protide1</b>	Numérique	Apport en protéines en % de l'apport total	sans unité	1819	7.1 à 17	quantitatif	
44	<b>lipide1</b>	Numérique	Apport en lipides en % de l'apport total	sans unité	1819	18.0 à 47.4	quantitatif	
45	<b>typeloge</b>	Numérique	Type de logement	1 : villa 2 : appartement 3 : studio 4 : maison arabe 5 : gourbi 6 : autre	1771	1 à 6	qualitatif	
46	<b>mode</b>	Numérique	Mode d'occupation du logement	1 : propriétaire 2 : locataire 3 : logé à titre gratuit 4 : accédant à la propriété	1774	1 à 4	qualitatif	
47	<b>nbchamb</b>	Numérique	Nombre de chambres du logement	Nombre de chambres	1602	1 à 9	quantitatif	
48	<b>bain</b>	Numérique	Salle de bain dans logement	1 : oui 2 : non	1820	1 à 2	dichotomique	
49	<b>douche</b>	Numérique	Douche dans logement	1 : oui 2 : non	1825	1 à 2	dichotomique	
50	<b>salleau</b>	Numérique	Salle d'eau dans logement	1 : oui 2 : non	1821	1 à 2	dichotomique	
51	<b>toilette</b>	Numérique	Toilettes dans logement	1 : oui 2 : non	1819	1 à 2	dichotomique	
52	<b>robinet1</b>	Numérique	Eau par robinet dans logement	1 : oui 2 : non	1823	1 à 2	dichotomique	
53	<b>puits</b>	Numérique	Eau à partir d'un puits	1 : oui 2 : non	1824	1 à 2	dichotomique	
54	<b>robinet2</b>	Numérique	Eau par robinet public	1 : oui 2 : non	1817	1 à 2	dichotomique	
55	<b>égout</b>	Numérique	Evacuation eaux usées par égout	1 : oui 2 : non	1818	1 à 2	dichotomique	
56	<b>fosse</b>	Numérique	Evacuation eaux usées par fosse sceptique	1 : oui 2 : non	1803	1 à 2	dichotomique	

Numéro d'ordre dans le fichier	Nom de la variable	Format	Contenu	Unités ou codes	N	Minimum - maximum	Type statistique	Remarques diverses
57	<b>electric</b>	Numérique	Logement raccordé au réseau électrique	1 :oui 2 :non	1828	1 à 2	dichotomique	
58	<b>sonede</b>	Numérique	Logement raccordé réseau sonede (eau)	1 :oui 2 :non	1821	1 à 2	dichotomique	
59	<b>frigo</b>	Numérique	Logement équipé d'un réfrigérateur	1 :oui 2 :non	1792	1 à 2	dichotomique	
60	<b>phone</b>	Numérique	Logement équipé d'un téléphone	1 :oui 2 :non	1810	1 à 2	dichotomique	
61	<b>tele</b>	Numérique	Logement équipé d'une télévision	1 :oui 2 :non	1817	1 à 2	dichotomique	
62	<b>video</b>	Numérique	Logement équipé d'un magnétoscope	1 :oui 2 :non	1823	1 à 2	dichotomique	
63	<b>cuisine</b>	Numérique	Logement équipé d'une cuisinière	1 :oui 2 :non	1824	1 à 2	dichotomique	
64	<b>lavess</b>	Numérique	Logement équipé d'un lave vaisselle	1 :oui 2 :non	1837	1 à 2	dichotomique	
65	<b>clime</b>	Numérique	Logement équipé d'un climatiseur	1 :oui 2 :non	1825	1 à 2	dichotomique	
66	<b>central</b>	Numérique	Logement équipé du chauffage central	1 :oui 2 :non	1835	1 à 2	dichotomique	
67	<b>parabole</b>	Numérique	Logement équipé d'une parabole	1 :oui 2 :non	1834	1 à 2	dichotomique	
68	<b>lavlinge</b>	Numérique	Logement équipé d'un lave linge	1 :oui 2 :non	1822	1 à 2	dichotomique	
69	<b>menage</b>	Numérique	Présence d'une aide ménagère	1 :oui 2 :non	1836	1 à 2	dichotomique	
70	<b>voiture</b>	Numérique	Ménage possède une voiture	1 :oui 2 :non	1822	1 à 2	dichotomique	
71	<b>confort</b>	Numérique	Score de confort matériel du ménage	Nombre d'éléments de confort matériel	1719		quantitatif	A partir des 12 variables de possessions de frigo à voiture
72	<b>confc3</b>	Numérique	Classes de score de confort du ménage	1: niveau de confort faible(0 à2) 2: niveau de confort moyen (3) 3: niveau de confort élevé (4à10)	1719		qualitatif	Recodage de confort
73	<b>eco1</b>	Numérique	Indice de niveau économique du ménage	Sans unité (classement des ménages par indice de niveau de vie croissant)	1798	0 à 100	quantitatif	Est la coordonnée du ménage sur le premier axe de l'AFC sur un ensemble de variables d'équipement et de niveau de vie du ménage. Utilise typologie, mode et les 23 variables de bain à voiture
74	<b>eco1c3</b>	Numérique	Terciles de niveau économique	1 : 1 <sup>er</sup> tercile « bas » 2 :2 <sup>ème</sup> tercile « moyen » 3 : 3 <sup>ème</sup> tercile « élevé »	1798	1 à 3	qualitatif	Recodage de eco1. Les terciles sont estimés sur l'ensemble des ménages.
75	<b>eco1c3b</b>	Numérique	Terciles de niveau économique par milieu (rural vs urbain)	1 : 1 <sup>er</sup> tercile « bas » 2 :2 <sup>ème</sup> tercile « moyen » 3 : 3 <sup>ème</sup> tercile « élevé »	1798	1 à 3	qualitatif	Recodage de eco1. Les terciles sont estimés séparément dans chaque milieu (rural vs urbain)
76	<b>depenses</b>	Numérique	Dépenses alimentaires mensuelles	Dinars courants à la date d'enquête	1524	20 à 650	quantitatif	Déclaratif
77	<b>depc3</b>	Numérique	Classes de dépenses alimentaires mensuelles	1 :<100 dinars 2 :100 à <200 dinars 3 :>=200 dinars	1524	1 à 3	qualitatif	Recodage de depenses





**Analyse de données d'enquêtes épidémiologiques  
Exercices pratiques avec EpiData Analysis**

**5. Texte des exercices pratiques**

*N.B. : dans les exercices suivants, pour la mise en œuvre des bonnes pratiques (documentation, traçabilité) il sera indispensable de travailler à partir d'un fichier de commandes (.pgm). En optimiser la lisibilité par la mise en page et la rédaction de commentaires adéquats.*

*Au-delà des indications données dans les exercices, utiliser la documentation Epidata Analysis fournie pour des précisions sur les différentes commande, options etc.*

## 1- Statistique descriptive – Cas d'une variable

On travaillera suivant le cas à partir du fichier menages2010\_2ana.rec ou menfem2010\_2ana.rec

### 1.1 Une variable quantitative

- Représenter la distribution (histogram avec ou sans option /pct) de différentes variables (poids, taille, imc, calories totales, triglycérides, hémoglobine, proxy de niveau économique du ménage). Commenter.
- Pour la variable taille, faire varier la largeur des classes et voir l'effet sur l'aspect de la représentation (histogram avec option /start=110 par exemple, puis faire des classes de largeur 20, 10, 5, 2.5, 1 cm (option /width=). Commenter.
- Comment obtenir le tableau de fréquence correspondant à l'histogramme de taille avec des classes de largeur 5 cm ? (recode avec option by puis freq). Comparer l'utilisation de la commande bar sur la variable recodée avec celle de l'histogramme en classes de 5 cm.
- Calculer les statistiques de base sur la variable taille (commande means). Vérifier la valeur du minimum et maximum obtenue (sort + browse). Vérifier la valeur de la médiane et des quantiles P25 et P75 (on pourra utiliser freq avec l'option / cum). Que représente le graphique qu'on obtient par la commande cdfplot ? Peut-on y retrouver graphiquement les valeurs de la médiane, de P25 et P75 ?
- Faire une représentation « boxplot » de la distribution de taille pour l'ensemble des sujets (retrouver la valeur de la médiane et des quantiles P25 et P75), puis par « habitat ».
- Vérifier le calcul de la moyenne donné par Epidata Analysis (à partir de la somme). Calculer une nouvelle variable tailecmoy donnant pour chaque sujet l'écart entre sa taille et la moyenne de la taille sur l'ensemble des sujets. Calculer une variable tailecmoy2 donnant le carré de l'écart précédent. Visualiser ces variables (browse idfem taille, tailecmoy, tailecmoy2). Comment retrouver la valeur de la variance de la variable taille (cf. résultat de la commande means ci-dessus) à partir d'un calcul sur tailecmoy2 (describe ou means) ?
- Calculer une variable ztaille contenant le z-score de taille (par rapport à la distribution observée). Visualiser les variables taille, tailecmoy, ztaille (browse). Commenter. Représenter un histogramme de ztaille : interpréter l'axe des x.

### 1.2 Une variable qualitative

- Donner la distribution de fréquence de la variable statut matrimonial de la femme, *matc3* (freq). Voir l'effet des options /c, /cum, /m – ensemble ou séparément) Les % cumulés ont-il un sens pour cette variable ? Représenter cette distribution (bar ou pie). Les calculs obtenus par describe *matc3* ont-il un sens ?
- Donner la distribution de fréquence de la variable imc en 4 classes *imcc4* (freq). Voir l'effet des options /c, /cum, /m – ensemble ou séparément) Les % cumulés ont-il un sens pour cette variable ? Représenter cette distribution (bar ou pie). Les calculs obtenus par describe *imcc4* ont-il un sens ?
- Comparer le résultat des commandes freq *matc3* et tables *matc3*. Comparer le résultat de : freq *matc3 imcc4* avec celui de tables *matc3 imcc4*?
- Donner la distribution de fréquence de la variable binaire *obesite* (freq). Appliquer la commande describe à la variable *obesite*. Comparer la fréquence d'obésité obtenue avec freq avec le calcul de la moyenne : commentaire ?
- Générer trois variables 0/1 *matc31*, *matc32*, *matc33* codant les 3 modalités de *matc3* (avec recode). Comparer la moyenne (describe) de ces 3 nouvelles variables avec la distribution de fréquence de *matc3* obtenue ci-dessus.

## 2 – Variables aléatoires - Distribution d'échantillonnage - Intervalle de confiance

- Représenter (histogram) la distribution des variables taille puis imc. Commenter la forme des distributions de ces deux variables (notamment par rapport à la loi normale). Confirmer avec la commande `cdfplot` et option `/p` (droite de normalité). Générer une nouvelle variable `logimc=log(imc)`. Représenter sa distribution avec un histogramme puis une droite de normalité. Commentaire ?
- Appliquer la commande `means` à la variable `imc`. Identifier les différentes statistiques fournies : en particulier, vérifier le calcul de l'écart-type d'échantillonnage (`std err`) ainsi que celui de l'intervalle de confiance (niveau 0,95 par défaut). A partir des éléments fournis, comment calculer vous-même un intervalle de confiance de niveau 0,99 ?
- Appliquer la commande `means` à la variable binaire 0/1 `obesite`. Vérifier le calcul de la variance, de l'écart-type, de l'écart-type d'échantillonnage (`std err`) ainsi que de l'intervalle de confiance.
- Utiliser `freq` avec option `/ci` sur la variable obésité pour retrouver les résultats ci-dessus.
- Donner un intervalle de confiance pour les proportions des différents niveaux scolaires, variable `scoc4` (`freq`)
- Représenter graphiquement un intervalle de confiance pour la proportion de femmes obèses dans chaque région (`ciplot`).

## 3 – Premiers exemples de tests

### 3.1 Comparaison d'une moyenne à une constante

On s'intéresse à la variable donnant l'ingéré énergétique rapporté aux besoins (`calp`).

- Représenter la distribution de cette variable (histogram). Calculer les statistiques descriptives (`means`). Commenter.
- On s'intéresse à tester l'hypothèse que dans la population d'étude la moyenne de `calp=1`. Générer (`gen`) une variable `calptest=calp-1`. Représenter la distribution de cette variable (histogram). Calculer les statistiques descriptives (`means`). Commenter.
- En utilisant l'option `/t` de `means`, tester (au risque  $\alpha=0,05$ ) l'hypothèse que dans la population la moyenne de `calptest` est égale à zéro. Que concluez vous par rapport à l'hypothèse initiale (moyenne de `calp=1` dans la population) ? Cette conclusion est-elle conservée si on prend un risque  $\alpha$  de 0,01 ? Un risque  $\alpha$  de 0,001 ?

### 3.2 Comparaison de plusieurs moyennes (1 variable quantitative x 1 variable qualitative)

On désire comparer la distribution de `calp` entre les 4 types de milieu (habitat).

- Faire une représentation « boxplot » de la distribution de `calp` par habitat. Commenter.
- Calculer les statistiques descriptives de `calp` par habitat (`means` avec option `/by=`). Commenter
- Au risque  $\alpha=0,05$  tester l'hypothèse d'égalité des moyennes de `calp` dans les 4 types de milieux (`means` avec option `/by=` et `/t`). Conclusion ? Cette conclusion est-elle conservée si on prend un risque  $\alpha$  de 0,01 ?

## 4 - Deux variables qualitatives

### 4.1 Deux variables qualitatives (tableau 2 x 2)

- Construire le tableau d'effectifs correspondant au croisement *milieu x activité professionnelle (milieu x actpc2)* avec milieu en lignes et activité professionnelle en colonnes (tables) : vérifier la cohérence de l'effectif total avec les données manquantes pour chacune des variables ; vérifier les totaux marginaux.
- Faire ensuite afficher les % en colonnes (option /c). Puis ensuite les % en lignes (option /l). Vérifier les calculs des %.
- Parmi les femmes ayant une activité professionnelle, quel % se trouve en milieu rural ? Parmi les femmes du milieu urbain, quel % ont une activité professionnelle ? Et parmi-celles du milieu rural ?
- Faire afficher la valeur de la statistique de test chi-deux (option /t). Vérifier le calcul de la statistique de test : construire d'abord le tableau théorique puis calculer le chi-deux observé. Peut-on considérer que dans la population d'étude, le % de femmes ayant une activité professionnelle est différent en milieu urbain vs rural ?

### 4.2 Deux variables qualitatives (tableau l x c)

- Construire le tableau croisé *habitat x scoc3* avec habitat en lignes et niveau scolaire en colonnes en affichant les % lignes (répartition du niveau scolaire dans chaque type d'habitat) et le test du  $\chi^2$  d'indépendance (tables options /r et /t).
- Représenter graphiquement les % lignes (select et bar ou bar avec option if).
- Commentez les résultats.

## 5 - Deux variables qualitatives, indices d'association épidémiologiques

On étudie la relation entre obésité et milieu de résidence (milieu en deux classes ou habitat en 4 classes).

### 5.1 Cas des tableaux 2 x 2 (variable « maladie » dichotomique, variable « exposition » dichotomique)

On s'intéresse à la relation entre milieu de résidence urbain vs rural et obésité.

- Créer une variable *urbain* codant les femmes résidant en milieu urbain 1 : oui 0 :non ; e.g. syntaxe `gen urbain=(milieu=1)`.
- Construire le tableau croisé *urbain x obese* (tables) avec urbain en lignes et obese en colonnes, en faisant afficher les % d'obese par milieu urbain vs not (% lignes, option /r) ; quels sont les % d'obésité dans chacun des milieux ? Le tableau 2 x 2 est-il dans la présentation « habituelle » des tableaux 2 x 2 en épidémiologie ?
- Représenter graphiquement ces pourcentages d'obésité par milieu (ciplot). Comment obtenir les valeurs numériques des intervalles de confiance représentés sur le diagramme (freq) ?
- Demander le calcul du test du chi-deux ; commenter le résultat, i.e. dans la population d'étude le % d'obésité est-il différent par milieu ?
- Compléter l'analyse en demandant le calcul de l'Odds-ratio (OR) et du Risque Relatif (RR) quantifiant la modification du risque d'obésité en milieu urbain vs milieu rural (options /rr et /o). Attention également à l'ordre des catégories : sont-elles dans l'ordre habituel pour interpréter correctement les indices d'association. Vérifier les calculs des valeurs de l'OR et du RR ?
- Interpréter les intervalles de confiance de l'OR et du RR : sont-ils cohérents avec les analyses précédentes ?
- Créer une variable 0/1 codant le milieu rural. Calculer les OR et RR mesurant la modification du risque d'obésité quand on est en milieu rural vs urbain (tables avec options /rr et /o). Sont-ils cohérents avec les calculs précédents ?

### 5.2 Cas des tableaux 2 x l (variable « maladie » dichotomique, variable « exposition » >=2 modalités)

On s'intéresse à la relation entre milieu de résidence à 4 modalités (*habitat*) et obésité (*obese*).

- Construire le tableau croisé *habitat x obese* (habitat en lignes, obese en colonnes) et afficher les % d'obèses dans chaque type d'habitat et le test du chi-deux. Commentaire ?
- Représenter graphiquement le % d'obèses dans chaque habitat (ciplot).
- On désire compléter l'analyse en calculant pour chacune des 3 premières modalités de habitat, les 3 ORs et/ou 3 RRs mesurant la modification du risque d'obésité, par rapport à la dernière i.e. milieu rural dispersé (habitat=4) qui sera donc prise comme catégorie de référence (select et tables ou bien tables avec option if). Faire attention à l'ordre des catégories pour le calcul des OR et RR.

## 6- Deux variables quantitatives

- Représenter les graphiques bivariés (scatter) pour le poids en fonction de la taille; de même pour imc en fonction du tour de taille (tt) de même pour imc en fonction de l'ingéré calorique rapporté aux besoins (calp) ; de même pour les dépenses alimentaires (depenses) en fonction du proxy de niveau économique (eco1). Commenter/comparer les graphiques obtenus.
- Calculer les coefficients de corrélation linéaire correspondant à chacun de ces graphiques (correlate) et les interpréter en regard des graphiques précédents.
- Calculer la matrice de corrélation entre toutes les variables anthropométriques 2 à 2 ainsi que l'âge: age poids taille imc tt th rth. Par exemple pour poids x taille ou imc x tt, comparer les corrélations obtenues dans la matrice et celles obtenues précédemment : pourquoi sont-elles légèrement différentes ? Commenter les corrélations observées entre les 7 variables étudiées.

## 7 - Facteur de confusion – Ajustement – Facteur modificateur d'effet

On s'intéresse à l'effet de l'âge de la femme (*agec4*) sur la relation entre milieu de résidence en deux classes (*urbain*) et obésité (*obesite*).

- Construire le tableau croisé avec calcul de l'OR et du RR *urbain x obesite* (OR et RR bruts)
- Analyse stratifiée par âge : construire le tableau croisé et calculer l'OR et le RR *urbain x obesite* dans chacune des modalités de *agec4* (select et tables ou tables avec option if). Commenter : *agec4* est-il plutôt un facteur de confusion ou un facteur modificateur d'effet ?
- Obtenir de même une analyse stratifiée par âge de la relation *urbain x obesite* avec une seule instruction tables. Comparer avec les résultats obtenus ci-dessus.

De façon analogue s'intéresse à l'effet du niveau économique du ménage (*eco1c3*) sur la relation entre milieu de résidence en deux classes (*urbain*) et obésité (*obe*).

- Obtenir directement une analyse stratifiée par niveau économique de la relation *urbain x obesite* avec une seule instruction tables. Comparer les OR et RR bruts et ajustés. Commenter.