

```

//*****
// PROGRAMME : TDANACORUS 2010.pgm (pgm Epidata Analysis)
// DATE : début le 15/04/2010
// AUTEUR : P. Traissac, IRD, UMR 204 NUTRIPASS, pierre.traissac@ird.fr
// LIEU : Bureau IRD Montpellier
// BUT : Programme récapitulatif de tous les exercices d'analyse
//       Ecole thématique CORUS 2010 Rabat
//
//      DONNEES EN ENTREE : - menages2010_2.rec données menages
//                          avec nouvelles variables
//                          u.s. = menage (n=1763)
//                          - menfem2010_2.rec données femmes + ménages
//                          avec nouvelles variables
//                          u.s. = femme (n=1849)
//
//      EN SORTIE : divers graphiques et résultats
//*****

// tableau des données femmes
read "D:\corus\maroc\FormationRabat2010\data\menfem2010_2ANA.rec"

//*****
// 1 STATISTIQUE DESCRIPTIVE - 1 VARIABLE
//*****

// ***** 1 VARIABLE QUANTITATIVE *****

// on regarde différentes variables ( commande histogram "simple")
// effet de /pct (% au lieu d'effectifs)
histogram taille
// distributions symétriques unimodales
// de plus des points "outliers" pour cal
histogram taille /pct
histogram cal /pct

// distributions (un peu) dissymétriques unimodales
histogram poids /pct
histogram imc /pct
histogram trigly /pct
histogram hemoglob /pct

// distribution bimodale
histogram ecol

// effet de la largeur des intervalles de taille sur aspect de l'histogramme
histogram taille /start=110 /width=20 /pct
histogram taille /start=110 /width=10 /pct
histogram taille /start=110 /width=5 /pct
histogram taille /start=110 /width=2.5 /pct
histogram taille /start=110 /width=1 /pct

// pour obtenir le tableau de fréquence par exemple
// pour les classes de 5 cm, il faut créer une nouvelle variable
define tailcla5 ###
recode taille to tailcla5 by 5
freq tailcla5 /c /cum

```

```

// ces deux commandes donnent un résultat très proche
histogram taille /start=110 /width=5 /pct
bar tailcla5 /pct

// statistiques descriptives sur taille
means taille
describe taille
// vérifier valeur de min et max
sort taille
//browse
// vérifier la valeur des percentiles donnés par describe (median etc)
freq taille /c /cum
// on fait un cumulative plot sur lequel on peut retrouver p25, P50, P75 etc.
cdfplot taille
// représentation boxplot
// on retrouve médiane, p25 et p75
boxplot taille
boxplot taille / by=habitat

// vérifier la valeur de la moyenne ( à partir de sum et du nbre
d'observations)

// calculer variable "écart à la moyenne"
gen tailecmoy=taille - 156.82
// on regarde taille et l'écart à la moyenne
//browse idfem taille tailecmoy
// calcul variable "écart au carré"
gen tailecmoy2 = (taille - 156.82)^2
// calcul de la somme
des tailecmoy2
// on vérifie que si on divise par 1836-1=1835, on retrouve la variance
// et si on prend racine on trouve l'écart-type. CQFD

// calcul z-score
gen ztaille= (taille - 156.82)/6.2
// on regarde taille tailecmoy ztaille
// histogramme de ztaille
histogram ztaille /pct

```

```

//*****
// 1 STATISTIQUE DESCRIPTIVE - 1 VARIABLE
//*****

// ***** 1 VARIABLE QUALITATIVE *****

// données manquantes exclues des calculs
freq matc3 /c /cum
// données manquantes incluses des calculs
freq matc3 /c /cum /m
// % cum pas de sens (pas d'ordre)
// graphiques
bar matc3 / pct
pie matc3
// matc3 n'est pas quantitative mais un codage
// 1,2,3 du statut matrimonial
// describe matc3

// données manquantes exclues des calculs
freq imcc4 /c /cum
// données manquantes incluses des calculs
freq imcc4 /c /cum /m
// % cum a un sens (ordre)
// graphiques
bar imcc4 / pct
pie imcc4
// imcc4 n'est pas quantitative mais un codage
// 1,2,3,4 de classes d'IMC
// describe imcc4

// différence : freq donne deux tableaux univariés
//          tables un tableau croisé (+tard interprétation)
freq   matc3 imcc4
tables matc3 imcc4

// obésité est 0: non /1: oui
// donc somme =nbre de 1 = nbre d'obèses
// moyenne= % de 1 = % d'obèses
freq obese /c
describe obese

// indicatrices de matc3
define matc31 #
define matc32 #
define matc33 #
recode matc3 to matc31 1=1 2,3=0
recode matc3 to matc32 2=1 1,3=0
recode matc3 to matc33 3=1 1,2=0

// on vérifie que les moyennes des indicatrices sont les %
freq matc3 /c
des matc31 matc32 matc33

```

```

//*****
// 2 VARIABLES ALEATOIRES - INTERVALLES DE CONFIANCE
//*****
// distribution taille (semble normale)
histogram taille /pct
// droite de normalité OK
cdfplot taille /p
// distribution IMC (dissymétrique)
histogram imc /pct
// droite de normalité pas OK
cdfplot imc /p

// logimc est plus proche d'une normale
gen logimc=log(imc)
histogram logimc /pct
cdfplot logimc /p

// Statistiques sur imc
// ecart-type d'échantillonnage = stderr
// I.C. à 0.95 par [moy-1,960 stderr ,moy+1,960 stderr]
// I.C. à 0.99 par [moy-2,576 stderr ,moy+2,576 stderr]
means imc

// statistiques sur obesite (variable 0/1)
// I.C. à 0.95 par [moy-1,960 stderr ,moy+1,960 stderr]
means obesite
// on retrouve les résultats
freq obesite /c /ci

// I.C. sur modalités de scoc4
freq scoc4 /c /ci

//*****
// 3 PREMIERS EXEMPLES DE TESTS
//*****

// 3.1 HYPOTHESE MU=CONSTANTE
histogram calp /pct
means calp
// valeur de référence est 1 (=100% des besoins)
gen calptest=calp-1
// idem ci-dessus mais valeur de référence=0
histogram calptest /pct
means calptest
// calcul P-Value (Hypothèse moyenne calptest=0 dans la population)
// (P<0,0000001) rejet hypothèse nulle au seuil 0,05 / 0,01 / 0,0001
// dans la population, la moyenne de calp est différente de 1
means calptest /t

// 3.2 COMPARAISON DE PLUSIEURS MOYENNES
// différence de tendance centrale (moyenne, médiane) pas énorme
boxplot calp /by=habitat
means calp /by=habitat
// calcul P-Value (hypothèse d'égalité des moyennes dans la population)
means calp /by=habitat/t
// P= 0,0115
// rejet de H0 au seuil 0,05 mais pas 0,01

```

```

//*****
// 4 DEUX VARIABLES QUALITATIVES
//*****

//***** 4.1 Tableau croisé 2x2 *****
// % lignes, colonnes en détail
tables actpc2 milieu
// % colonnes (c=columns)
tables actpc2 milieu/c
// % lignes (r=rows)
tables actpc2 milieu /r
// calcul test chi-deux + P-Value
// P-Value<0,00001 : rejet hypothèse nulle
// les % d'activité professionnelles dans la population
// sont différents par milieu
tables actpc2 milieu /r /t

//***** 4.1 Tableau croisé quelconque *****
// tableau croisé avec % lignes et test chi-deux
// les % de niveau scolaires dépendent de habitat
tables scoc3 habitat /r /t
// cela ne donne pas les bonnes représentations
// on ne retrouve pas mes % ligne du tableau croisé
bar scoc3 /pct / by=habitat
// cela donne les distributions de scoc3 par habitat
// idem % lignes du tableau croisé
bar scoc3 /pct if habitat=1
bar scoc3 /pct if habitat=2
bar scoc3 /pct if habitat=3
bar scoc3 /pct if habitat=4

```

```

//*****
// 5 DEUX VARIABLES QUALITATIVES
// INDICES D'ASSOCIATION EPIDEMIOLOGIQUES
//*****
//*** 5.1 Cas des tableaux 2 x 2 *****
//generation variable 0/1 codant urbain
// pas de précaution spéciale pour les valeurs
// manquantes car il n'y en a pas pour cette variable
gen urbain=(milieu=1)
// verification
freq milieu urbain
// tableau croisé "normal"
// présentation par valeurs croissantes
// des lignes et colonnes (ie pas celui "épidémiologique")
tables obesite urbain /r/t
// représentation des % obèses par milieu
ciplot obesite urbain
// intervalles de confiance des % d'obèses par milieu
freq obesite /c /ci if urbain=1
freq obesite /c /ci if urbain=0

// table avec indices épidémiologiques
// la presentation set met "automatiquement" par valeurs décroissantes
// des lignes et colonnes (inverse de précédemment)car demande OR et RR
tables obesite urbain /r/t /rr /o

// variable rural 0/1
gen rural=(milieu=2)
// Les indices sont inversés car on a inversé exposé / non exposé
tables obesite rural /r/t /rr /o

//*** 5.2 Cas des tableaux 2 x 1 *****
// Tableau croisé, % ligne, test du chi-deux
// le % d'obèse dépend de habitat (P<0,00001)
tables obesite habitat /r /t
// Graphique du %obésité + I.C. par habitat
ciplot obesite habitat

// Pour calcul des ORs on se ramène a des tableaux 2 x 2 par sélection
select habitat=1 or habitat=4
tables obesite habitat /r/t /rr /o
select
// est équivalent à
tables obesite habitat /r/t /rr /o if habitat=1 OR habitat=4

// mais les syntaxes if sont plus simples
// cela ne donne pas le bon ordre de catégories
tables obesite habitat /r/t /rr /o if habitat=1 OR habitat=4
tables obesite habitat /r/t /rr /o if habitat=2 OR habitat=4
tables obesite habitat /r/t /rr /o if habitat=3 OR habitat=4
// solution possible recoder habitat= 4 en habitat=0
if habitat=4 then habitat=0
tables obesite habitat /r/t /rr /o if habitat=1 OR habitat=0
tables obesite habitat /r/t /rr /o if habitat=2 OR habitat=0
tables obesite habitat /r/t /rr /o if habitat=3 OR habitat=0

```

```

//*****
// 6 DEUX VARIABLES QUANTITATIVES
//*****
// graphiques bivariés X x Y
scatter taille poids
scatter tt imc
scatter calp imc
scatter ecol dépenses

// NB : possible faire apparaitre "sous groupes"
scatter tt imc /by=milieu

// corrélation linéaires correspondant aux graphiques
correlate taille poids
correlate tt imc
correlate calp imc
correlate ecol dépenses

// matric de corrélation des variables anthropo
correlate age poids taille imc tt th rth

```

```

//*****
// 7 TROIS VARIABLES QUANTITATIVES
//   FACTEUR DE CONFUSION - AJUSTEMENT
//*****

// variable à créer si pas déjà fait ci-dessus
gen urbain=(milieu=1)
// analyse "brute"
tables  obesite urbain /r /t /o /rr

// Analyse stratifiée par âge « manuelle »
// Un peu long, de plus ne donne pas le RR ou OR ajusté
tables  obesite urbain /r /t /o /rr if agec4=20
tables  obesite urbain /r /t /o /rr if agec4=30
tables  obesite urbain /r /t /o /rr if agec4=40
tables  obesite urbain /r /t /o /rr if agec4=50

// Analyse stratifiée par âge avec 1 seul tables
// L'âge est plutôt un modificateur d'effet
// pour l'association obesite x milieu
// (modifie la force même si pas le sens de la relation)
// syntaxe : tables « maladie » « facteur » « facteur de confusion »
tables  obesite urbain agec4 /r /t /o /rr

// Analyse stratifiée par niveau économique
// On voit que +/- modificateur d'effet
// Mais on peut se poser aussi la question de l'ajustement
// cf association ajustée - forte qu'en brut
// Association entre obesite et urbain en partie due
// a differences socio-économiques
tables  obesite urbain ecolc3/r /t /o /rr

// si on veut ajuster simultanément sur plusieurs facteurs de confusion
// on les ajoute à la suite dans le tables
// mais attention aux problèmes d'effectifs dans les strates
// si trop de facteurs => trop de strates : analyses sur effectifs trop petits
// voire analyse impossible dans certains cas
// exemple analyse : obésité x urbain ajustée pour niveau éco en (3 classes)
// niveau scolaire (3 classes) et activité professionnelle (2 classes)
// ce qui résulte en 18 strates ... !
tables  obesite urbain ecolc3 scoc3 actpc2/r /t /o /rr

```