Projet Obe Maghreb

Ecole thématique gestion et analyse de données
20 au 29 avril 2010

Gestion et analyse de données d'enquêtes épidémiologiques

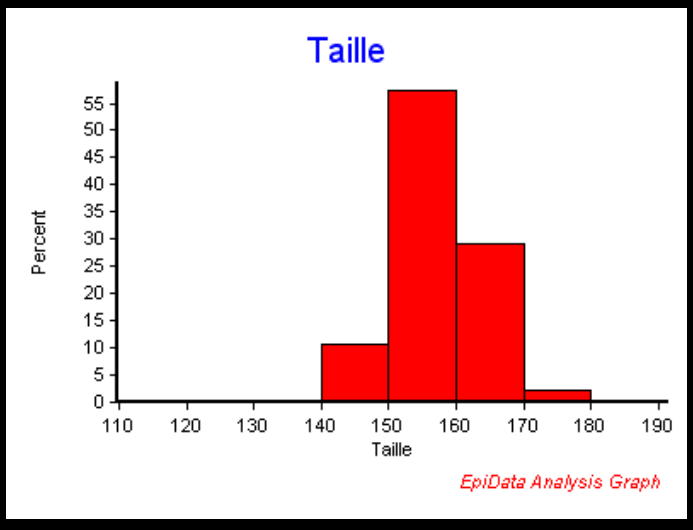
Analyse de données

1.Statististique descriptive
Cas d'une variable
Exemples utilisés pendant le cours

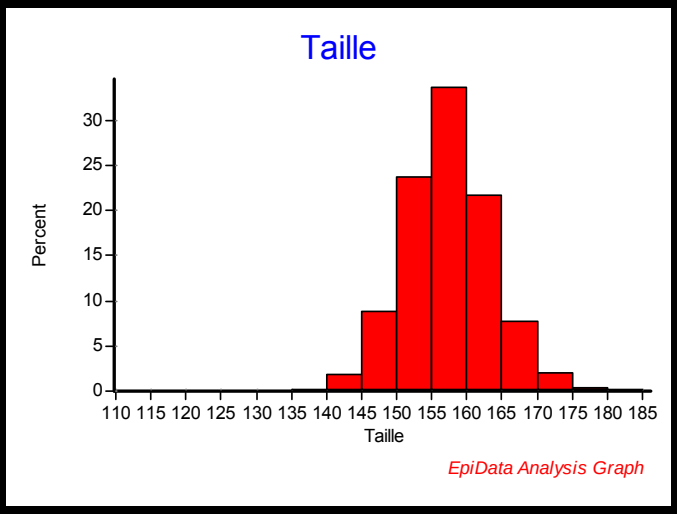
Pierre Traissac
UMR 204 « Prévention des malnutritions et pathologies associées »
IRD, Montpellier, France

Institut de recherche pour le développement

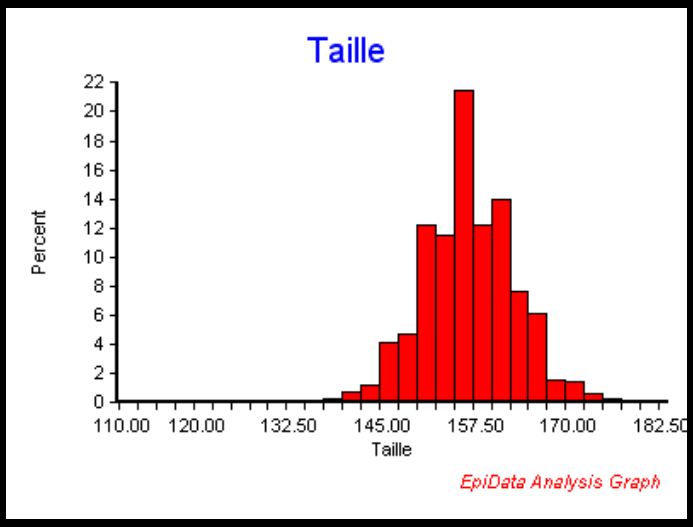
Histogramme de taille (n=1836)
Classes de 10 cm



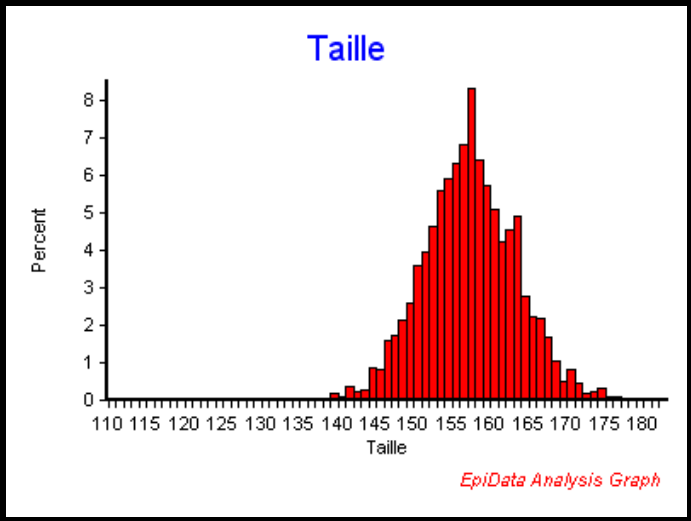
Histogramme de taille (n=1836)
Classes de 5 cm



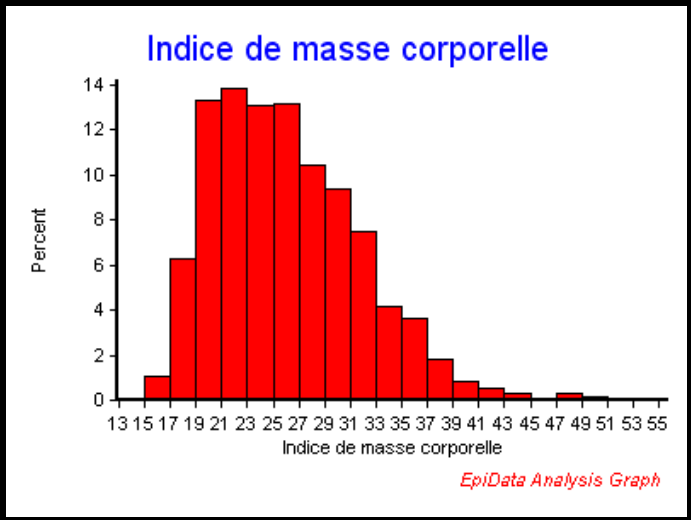
Histogramme de taille (n=1836)
Classes de 2,5 cm



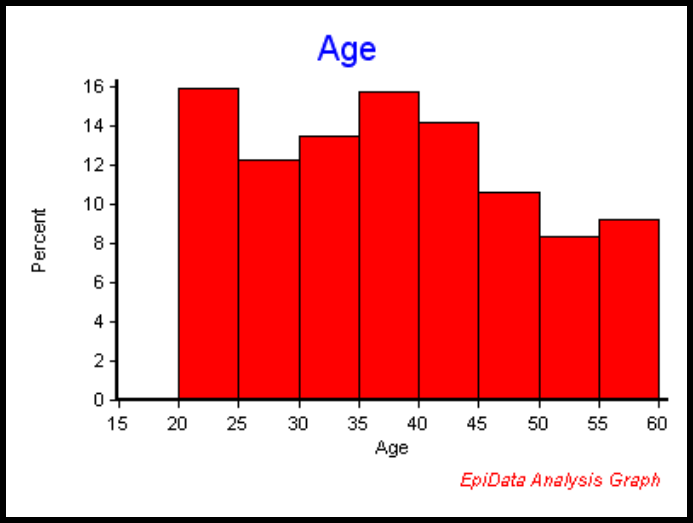
Histogramme de taille (n=1836)
Classes de 1 cm



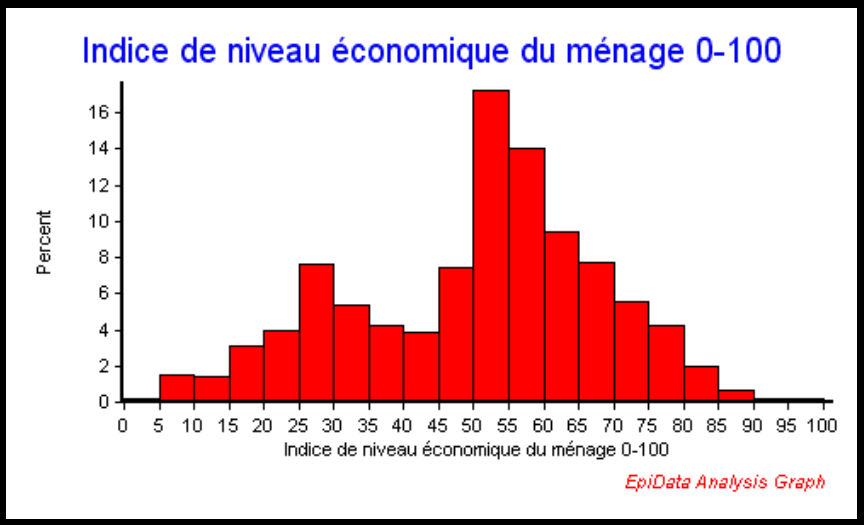
Histogramme d'IMC en kg/m^2 (n=1835)



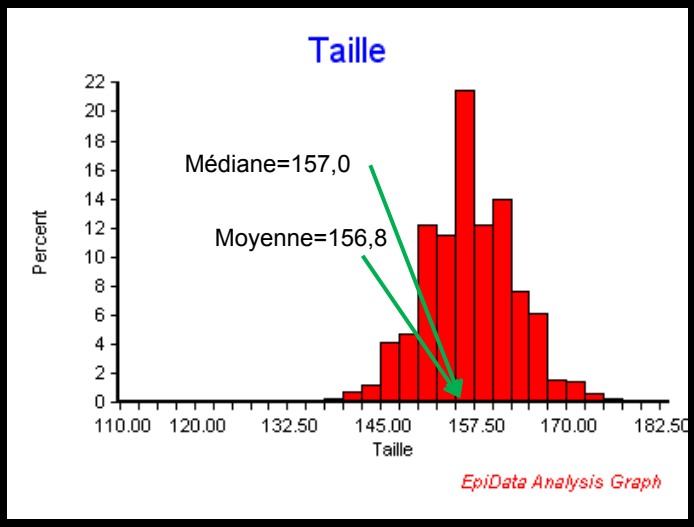
Histogramme d'age en années (n=1849)



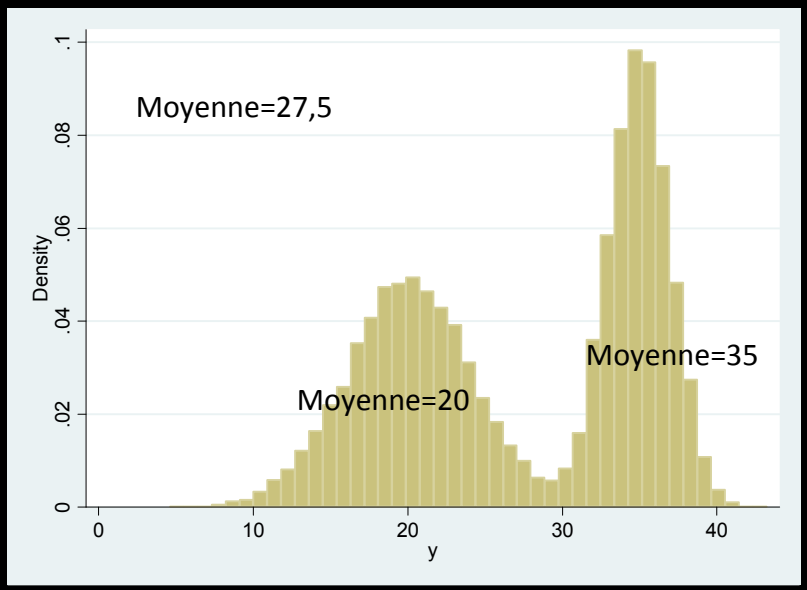
Histogramme proxy de niveau économique (n=1798)



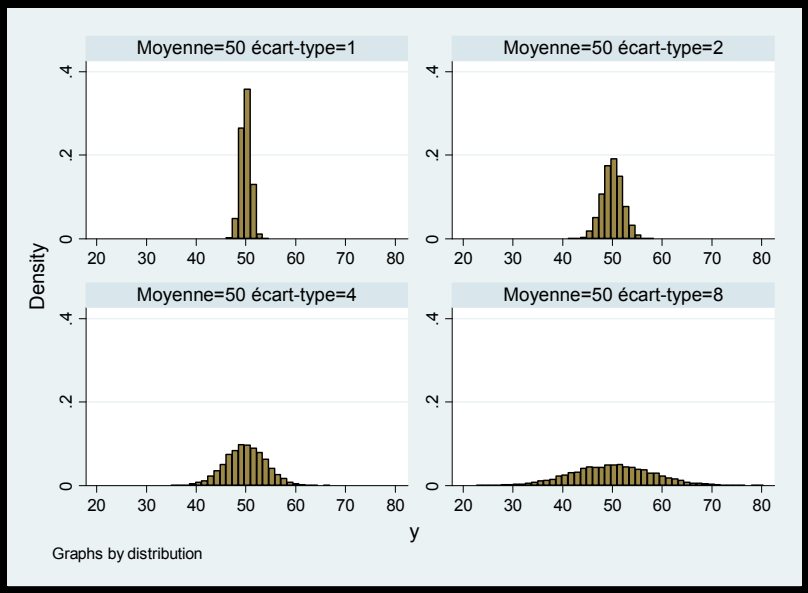
Tedance centrale : moyenne, médiane
Taille en cm (n=1836)



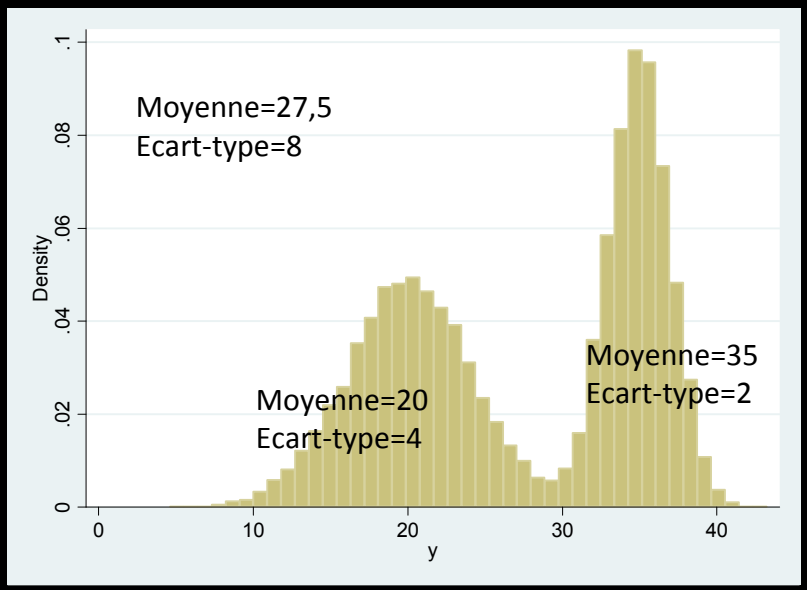
Moyenne/médiane distribution bi-modale ?



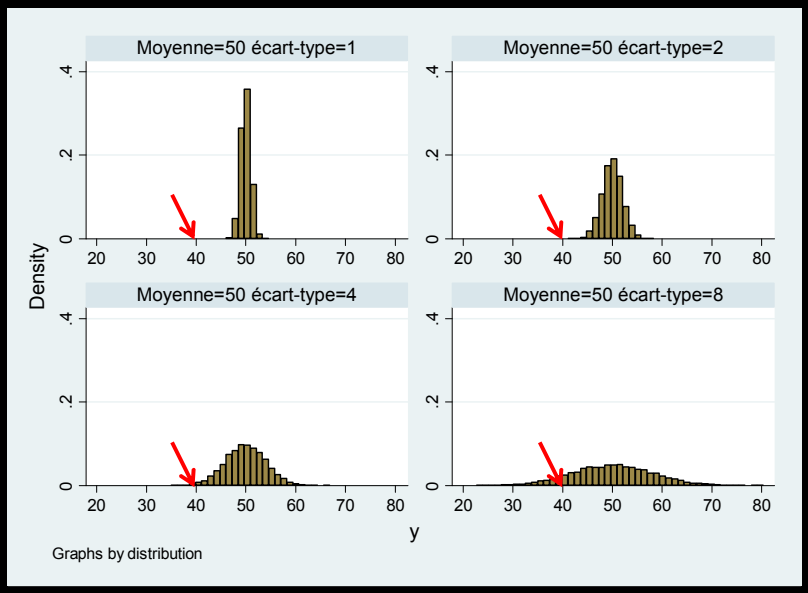
Dispersion : Variance – Ecart-type



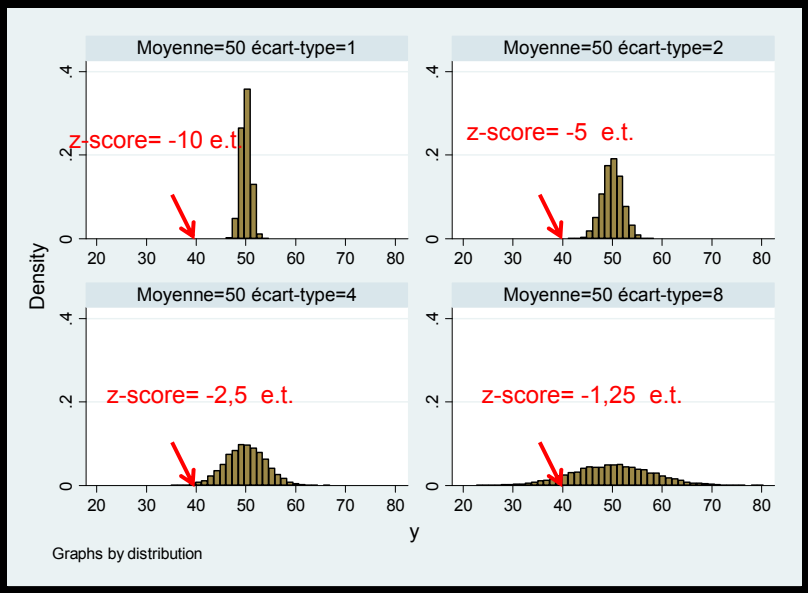
Variance / écart-type distribution bi-modale ?



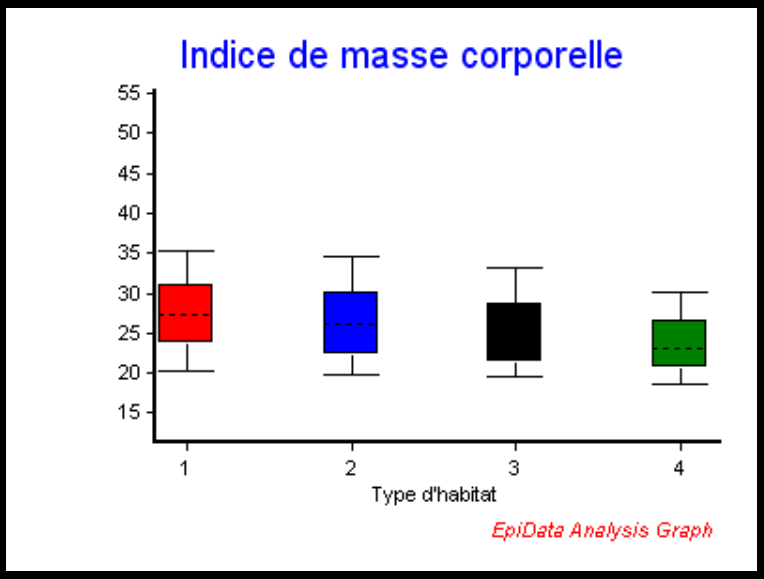
Z-score : moyenne=50, X=40



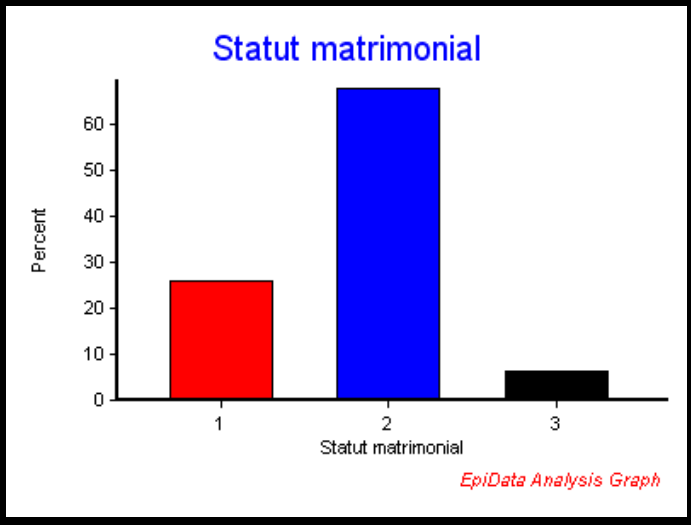
Z-score : moyenne=50, X=40



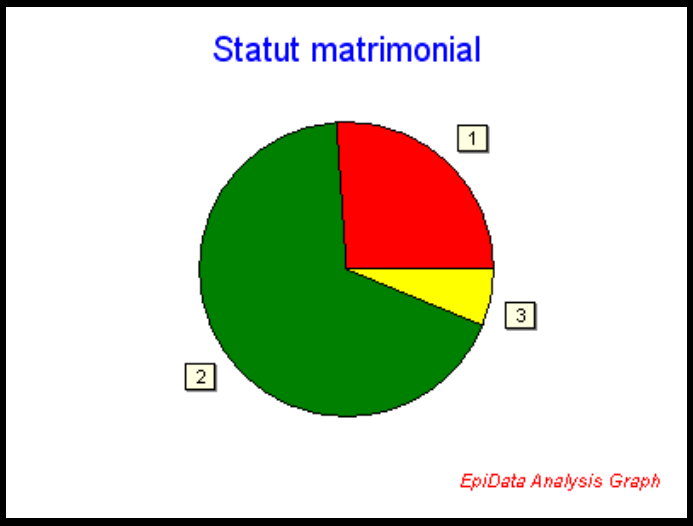
Box-plots IMC par habitat



Variable qualitative – Diagramme en barres
Statut matrimonial (n=1849)



Variable qualitative – Diagramme camembert
Statut matrimonial (n=1849)



Projet Obe Maghreb

Ecole thématique gestion et analyse de données
20 au 29 avril 2010

Gestion et analyse de données d'enquêtes épidémiologiques

Analyse de données

2. Probabilités, variables aléatoires
Distribution d'échantillonnage. Intervalle de confiance
Exemples utilisés pendant le cours

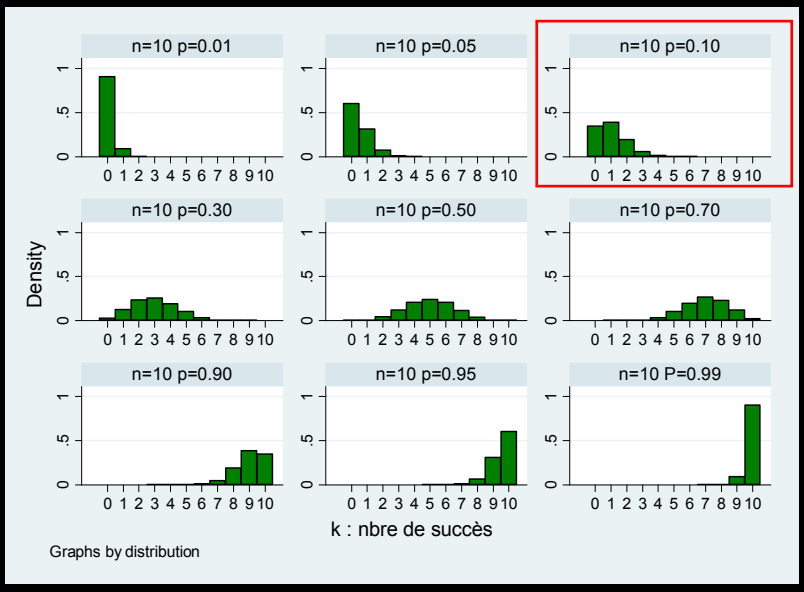
IRD
Institut de recherche pour le développement

Pierre Traissac
UMR 204 « Prévention des malnutritions et pathologies associées »
IRD, Montpellier, France

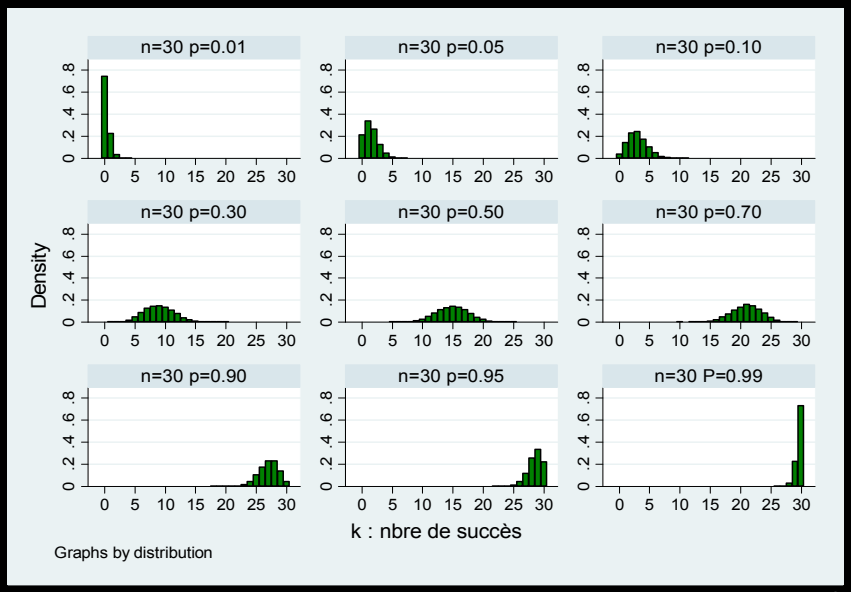
Exemples de lois binomiales n=10

K p	0,01	0,05	0,1	0,3	0,5	0,7	0,9	0,95	0,99
0	0,904	0,599	0,349	0,028	0,001	0,000	0,000	0,000	0,000
1	0,091	0,315	0,387	0,121	0,010	0,000	0,000	0,000	0,000
2	0,004	0,075	0,194	0,233	0,044	0,001	0,000	0,000	0,000
3	0,000	0,010	0,057	0,267	0,117	0,009	0,000	0,000	0,000
4	0,000	0,001	0,011	0,200	0,205	0,037	0,000	0,000	0,000
5	0,000	0,000	0,001	0,103	0,246	0,103	0,001	0,000	0,000
6	0,000	0,000	0,000	0,037	0,205	0,200	0,011	0,001	0,000
7	0,000	0,000	0,000	0,009	0,117	0,267	0,057	0,010	0,000
8	0,000	0,000	0,000	0,001	0,044	0,233	0,194	0,075	0,004
9	0,000	0,000	0,000	0,000	0,010	0,121	0,387	0,315	0,091
10	0,000	0,000	0,000	0,000	0,001	0,028	0,349	0,599	0,904

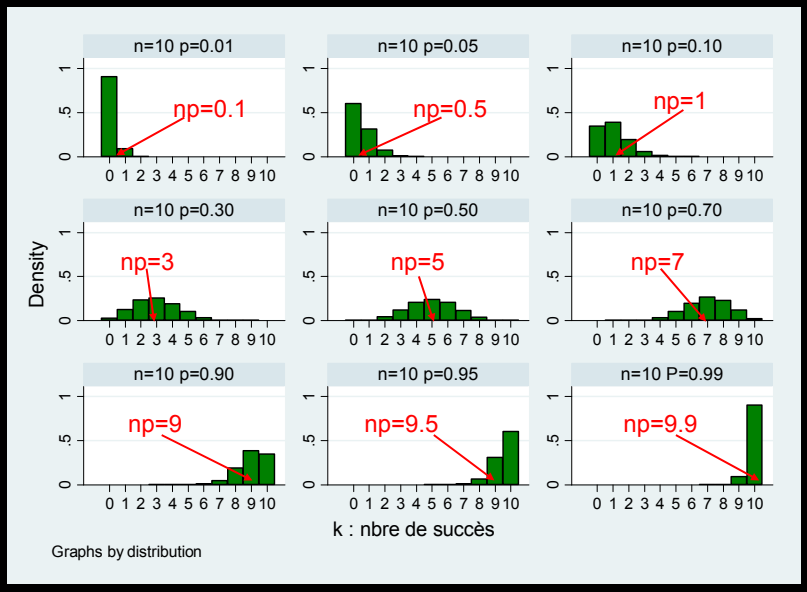
Exemples de loi binomiales n=10



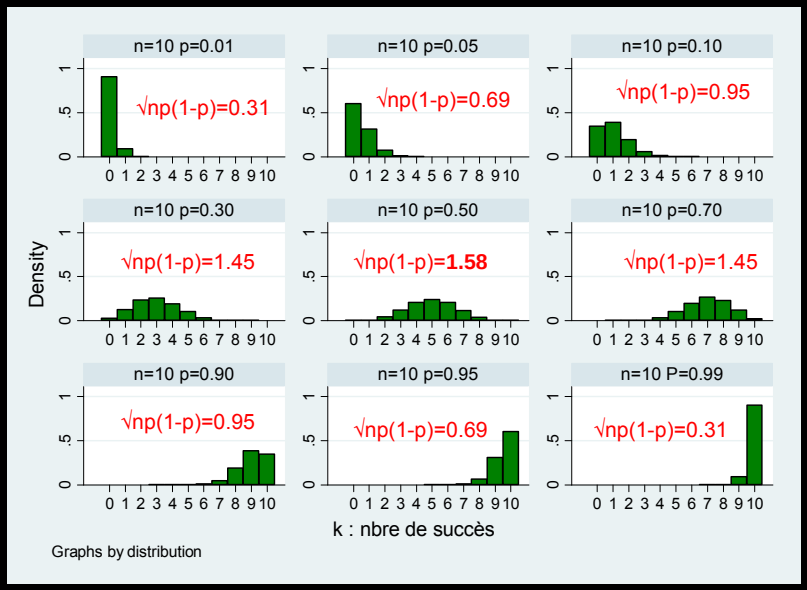
Exemples de loi binomiales n=30



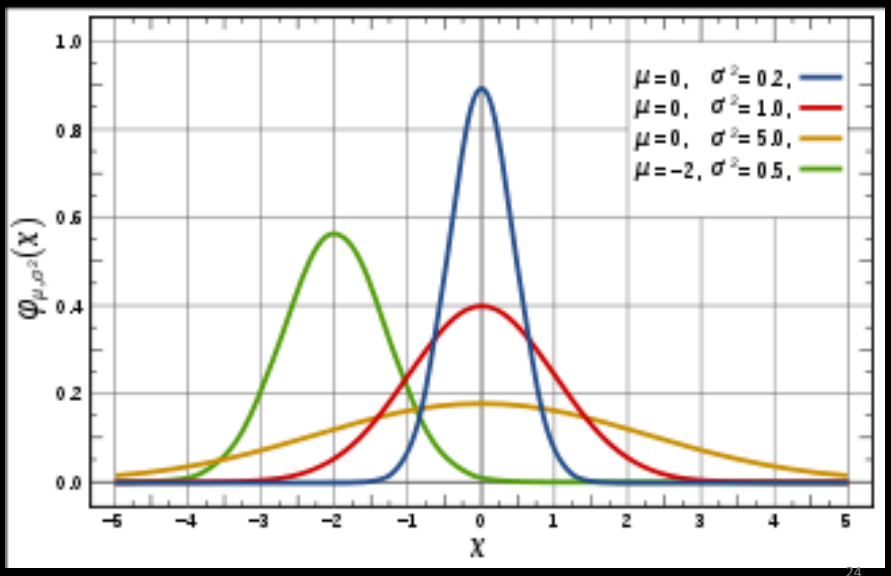
Espérance – binomiales n=10



Ecart-type - loi binomiales n=10



Loi normales

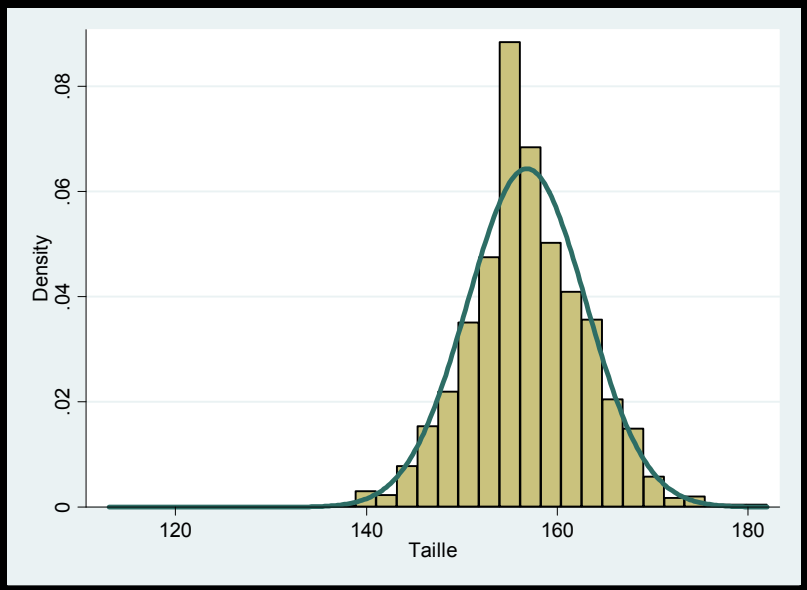


Loi normale N(0,1)

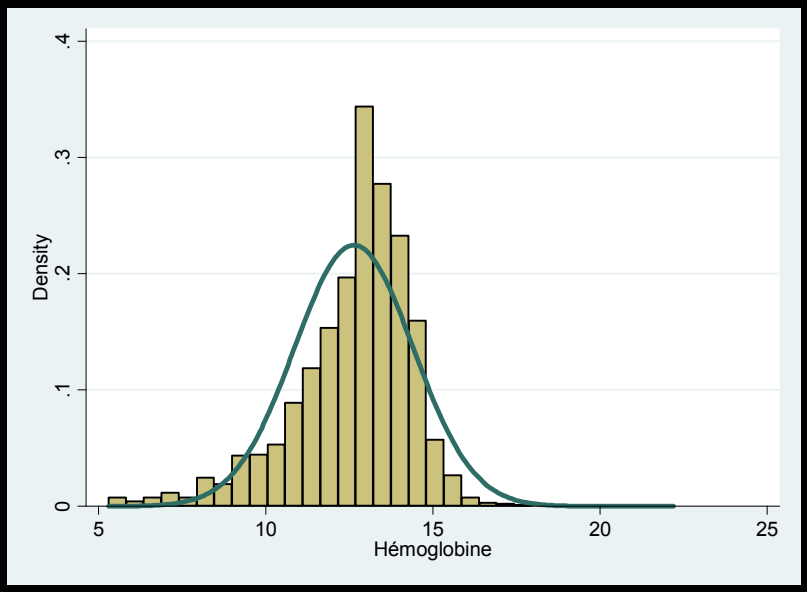
Probabilité d'être en dehors d'un intervalle symétrique autour de zéro
e.g. pour P=0,05 intervalle [-1,960;+1,960]
e.g. pour P=0,12 intervalle [-1,555;+1,555]

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00		2,576	2,326	2,170	2,054	1,960	1,881	1,812	1,751	1,695
0,10	1,645	1,598	1,555	1,514	1,476	1,440	1,405	1,372	1,341	1,311
0,20	1,282	1,254	1,227	1,200	1,175	1,150	1,126	1,103	1,080	1,058
0,30	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,860
0,40	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,690
0,50	0,674	0,659	0,643	0,628	0,613	0,598	0,583	0,568	0,553	0,539
0,60	0,524	0,510	0,496	0,482	0,468	0,454	0,440	0,426	0,412	0,399
0,70	0,385	0,372	0,358	0,345	0,332	0,319	0,305	0,292	0,279	0,266
0,80	0,253	0,240	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138
0,90	0,126	0,113	0,100	0,088	0,075	0,063	0,050	0,038	0,025	0,013

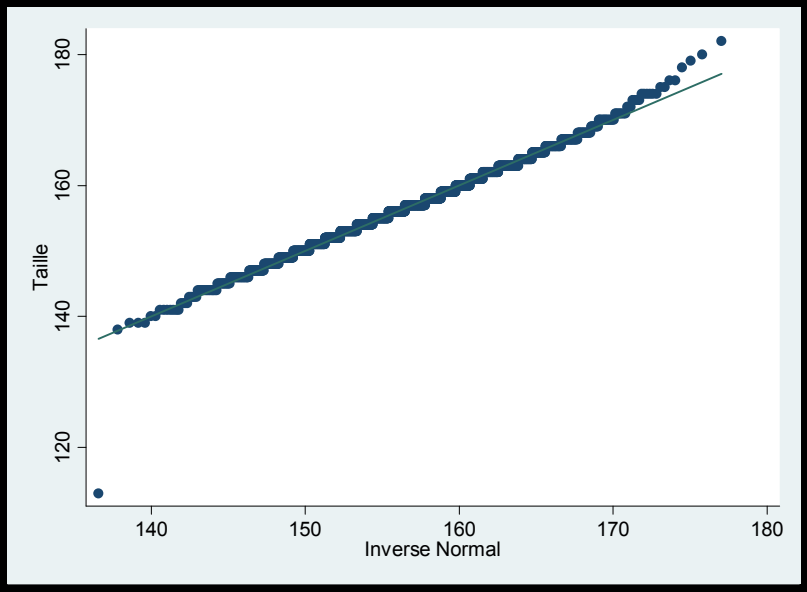
Distribution de taille en cm (n=1836)



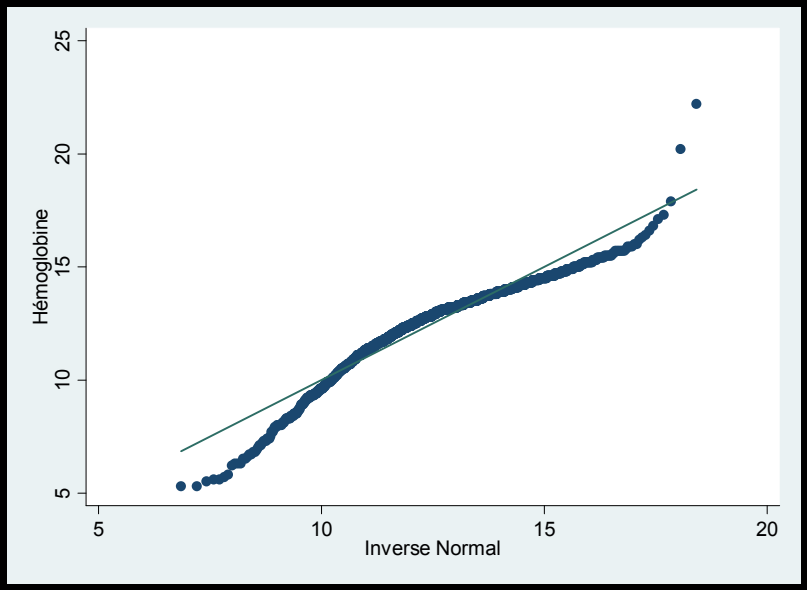
Distribution de hémoglobine en g/dl (n=1789)



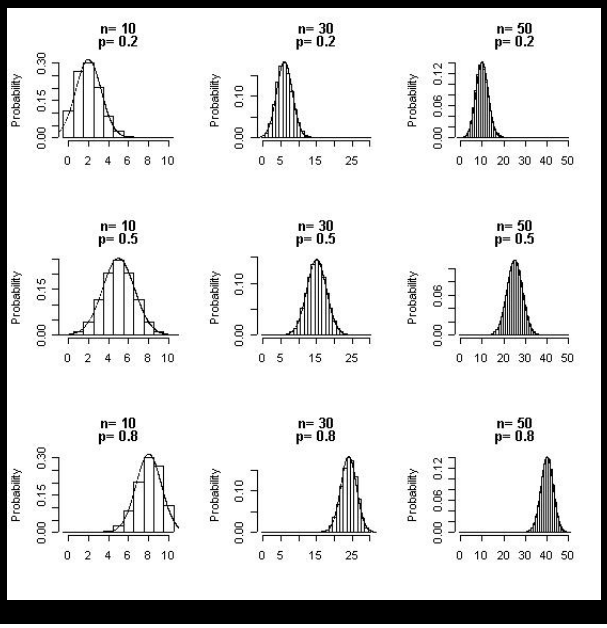
Taille : quantile plot (n=1836)



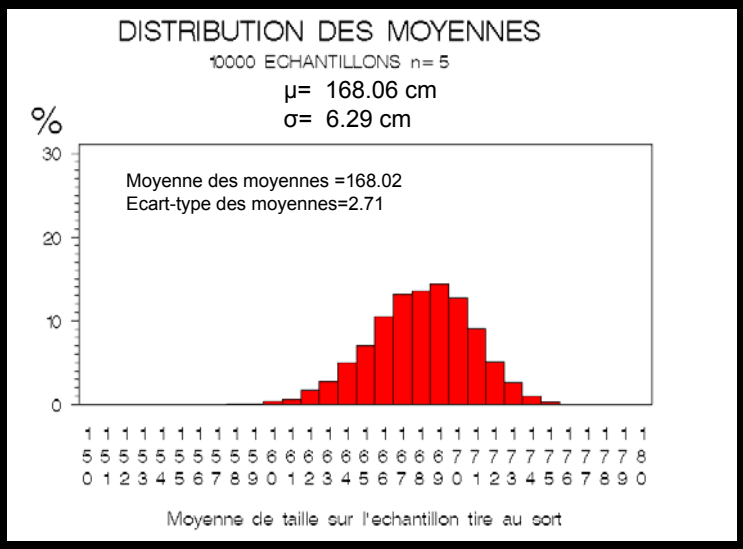
Hémoglobine : quantile plot (n=1789)



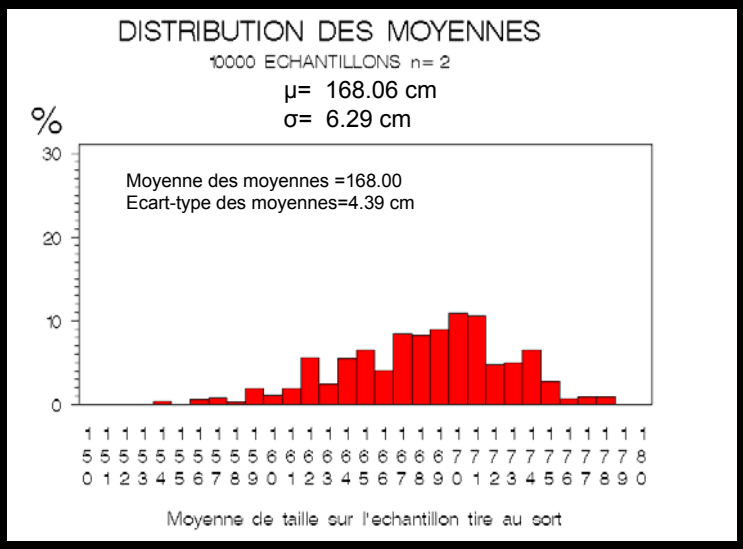
Approximation normale d'une loi binomiale



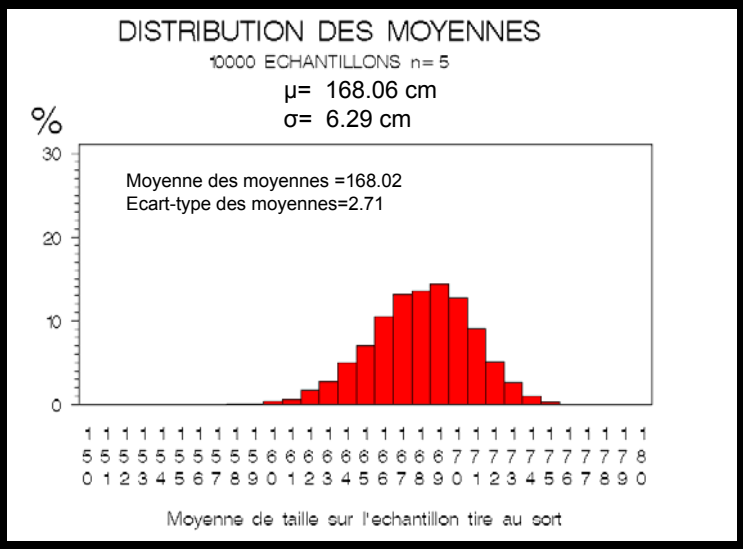
Simulation d'une distribution d'échantillonnage



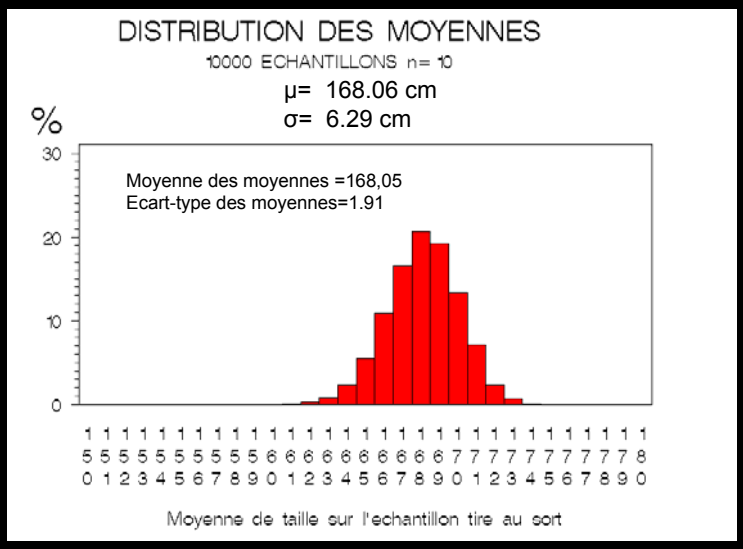
Simulation d'une distribution d'échantillonnage



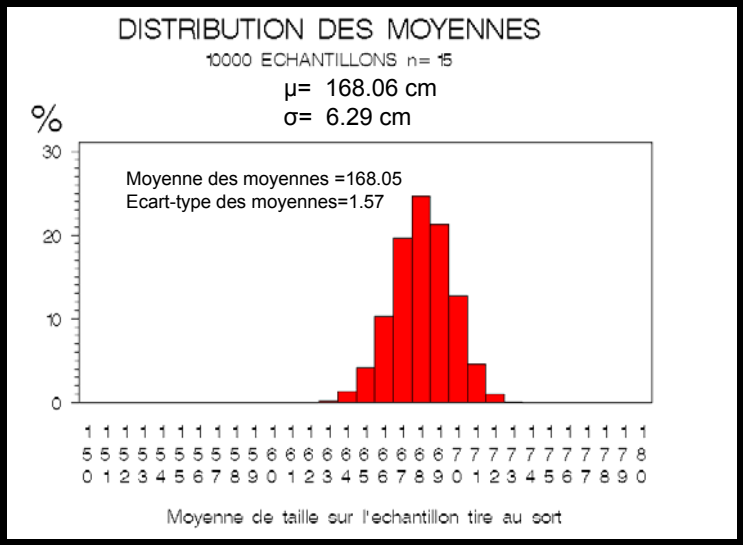
Simulation d'une distribution d'échantillonnage



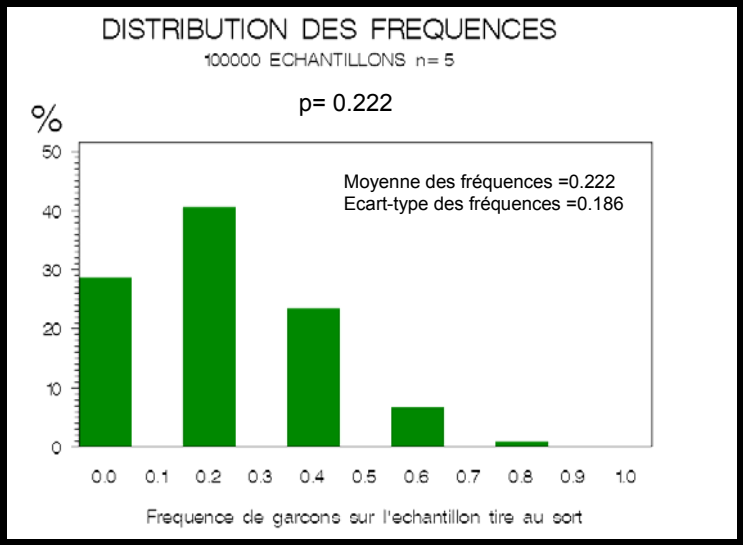
Simulation d'une distribution d'échantillonnage



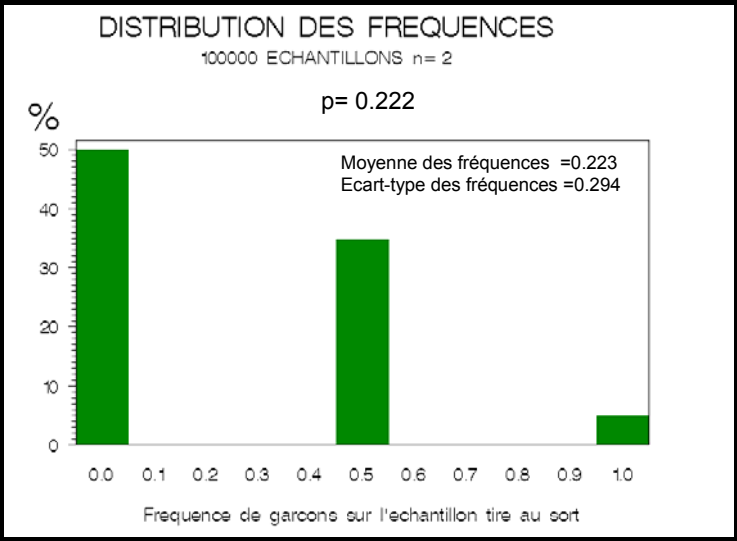
Simulation d'une distribution d'échantillonnage



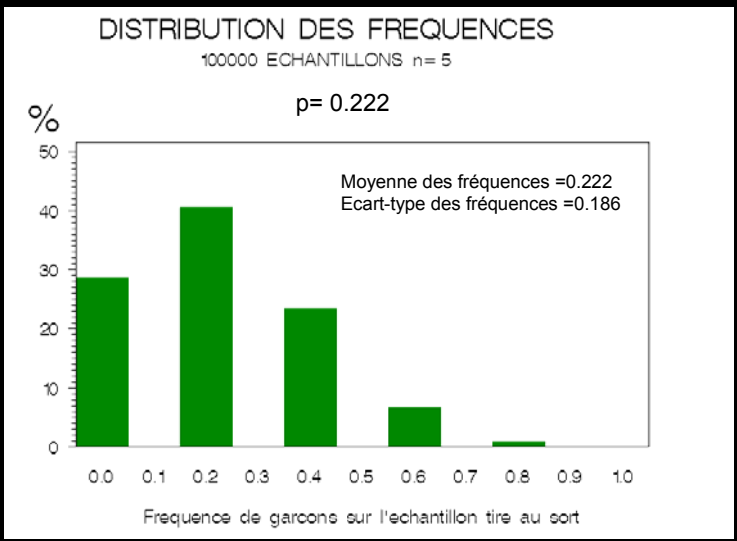
Simulation d'une distribution d'échantillonnage



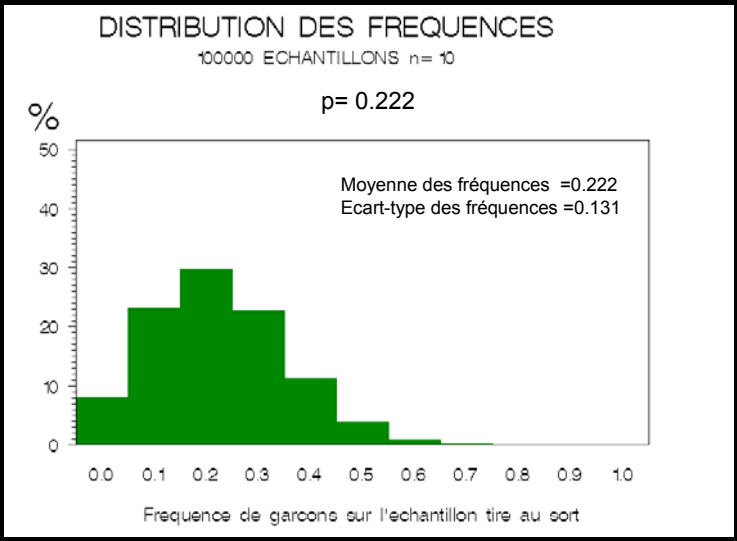
Simulation d'une distribution d'échantillonnage



Simulation d'une distribution d'échantillonnage



Simulation d'une distribution d'échantillonnage



CORUS

Projet Obe Maghreb

Ecole thématique gestion et analyse de données
20 au 29 avril 2010

Gestion et analyse de données d'enquêtes épidémiologiques

Analyse de données

4. Deux variables qualitatives
Tableaux croisés – Test du chi-deux
Exemples utilisés pendant le cours

IRD
Institut de recherche pour le développement

Pierre Traissac

UMR 204 « Prévention des malnutritions et pathologies associées »
IRD, Montpellier, France

Habitat x niveau économique (n=1725)
Effectifs observés

Habitat /Niveau économique	Bas	Moyen	Elevé	Total
Grandes villes	15	83	281	379
Autres villes	85	256	314	655
Rural	475	167	49	691
Total	575	506	644	1725

Habitat x niveau économique (n=1725)
% calculés en colonnes

Habitat /Niveau économique	Bas	Moyen	Elevé	Total
Grandes villes	2,6%	16,4%	43,6%	22,0%
Autres villes	14,8%	50,6%	48,8%	38,0%
Rural	82,6%	33,0%	7,6%	40,0%
Total	100,0%	100,0%	100,0%	100,0%

Habitat x niveau économique (n=1725)
% calculés en lignes

Habitat /Niveau économique	Bas	Moyen	Elevé	Total
Grandes villes	4,0%	21,9%	74,1%	100,0%
Autres villes	13,0%	39,1%	47,9%	100,0%
Rural	68,7%	24,2%	7,1%	100,0%
Total	33,3%	29,3%	37,4%	100,0%

Habitat x niveau économique (n=1725)
% lignes théoriques si indépendance
(i.e. pas d'association milieu x niveau éco)

Habitat /Niveau économique	Bas	Moyen	Elevé	Total
Grandes villes	33,3%	29,3%	37,4%	100,0%
Autres villes	33,3%	29,3%	37,4%	100,0%
Rural	33,3%	29,3%	37,4%	100,0%
Total	33,3%	29,3%	37,4%	100,0%

Habitat x niveau économique (n=1725)
Effectifs théoriques si indépendance (observés)

Habitat /Niveau économique	Bas	Moyen	Elevé	Total
Grandes villes	126,3 (15)	111,2 (83)	141,5 (281)	379
Autres villes	218,3 (85)	192,1 (256)	244,5 (314)	655
Rural	230,3 (475)	202,7 (167)	258,0 (49)	691
Total	575	506	644	1725

45

Habitat x niveau économique (n=1725)
Différence effectif observé - théorique

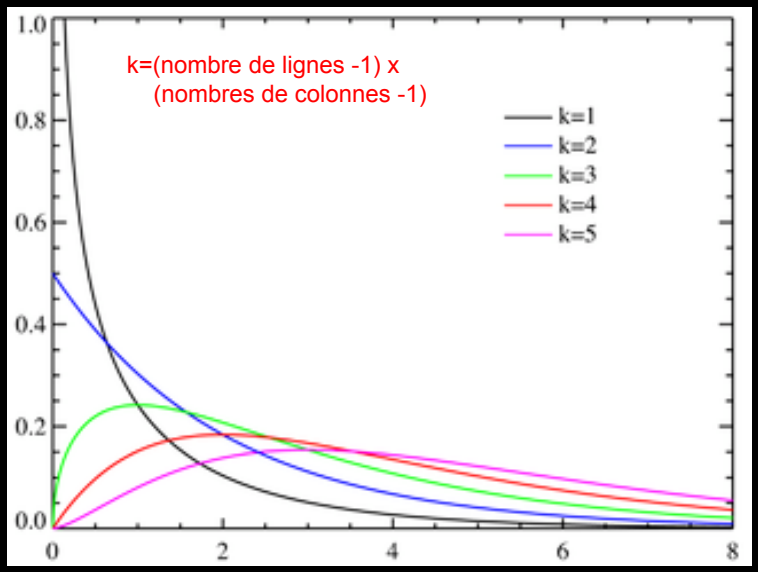
Habitat /Niveau économique	Bas	Moyen	Elevé	Total
Grandes villes	-111,3	-28,2	+139,5	0
Autres villes	-133,3	+63,9	+69,5	0
Rural	+244,7	-35,7	-209	0
Total	0	0	0	

46

Habitat x niveau économique (n=1725)
Contributions des cellules
à la valeur de la statistique chi-deux observée

Habitat /Niveau économique	Bas	Moyen	Elevé	Total
Grandes villes	98,1	7,1	137,5	
Autres villes	81,4	21,2	19,7	
Rural	259,9	6,3	169,3	
Total				800,6

Loi du chi-deux (i.e. de la statistique de test chi-deux
sous l'hypothèse nulle d'indépendance lignes x colonnes)



Projet Obe Maghreb

Ecole thématique gestion et analyse de données
20 au 29 avril 2010

Gestion et analyse de données d'enquêtes épidémiologiques

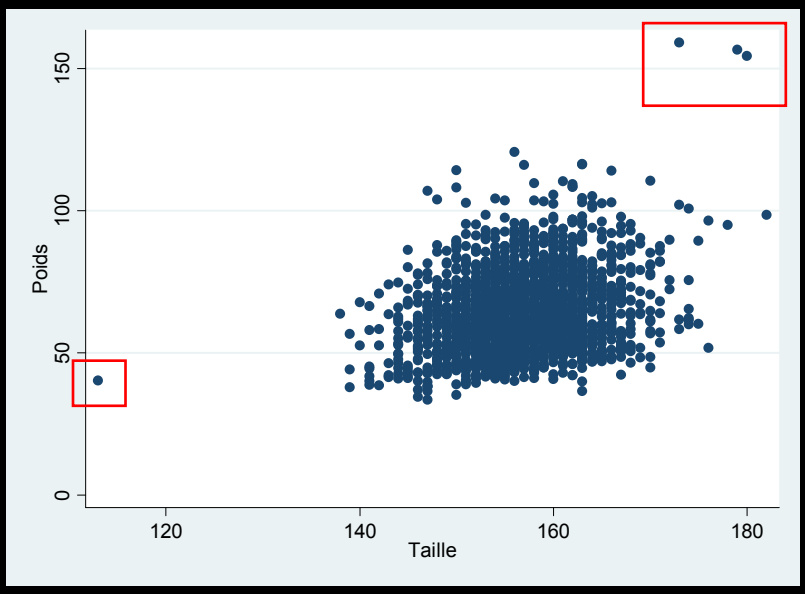
Analyse de données

6. Deux variables quantitatives
Exemples utilisés pendant le cours

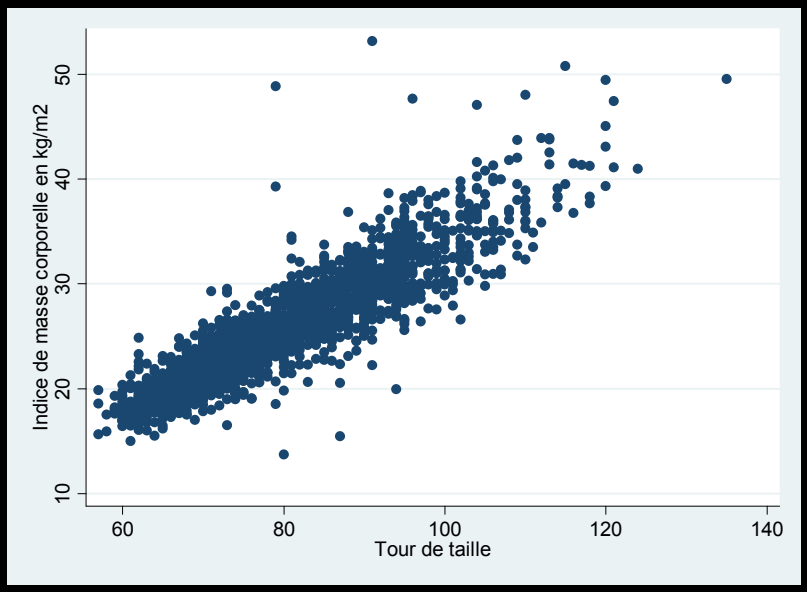
Pierre Traissac
UMR 204 « Prévention des malnutritions et pathologies associées »
IRD, Montpellier, France

Institut de recherche pour le développement

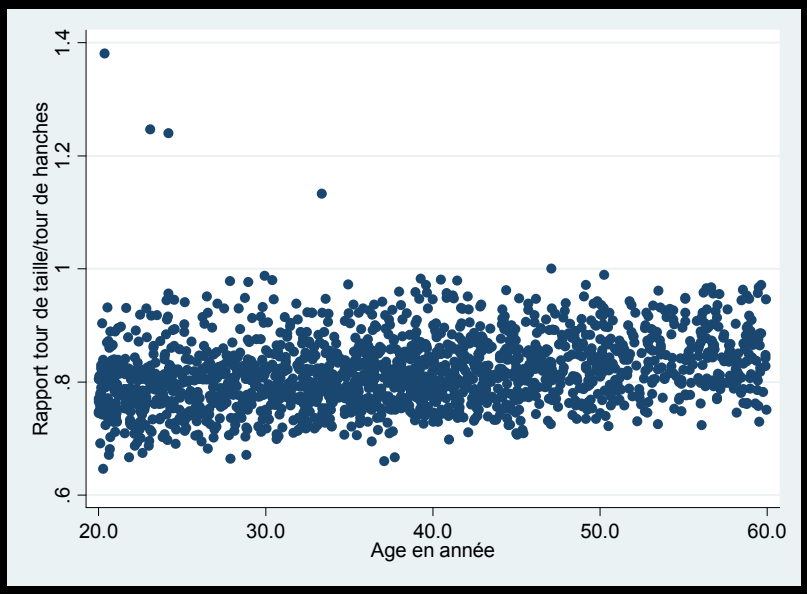
Poids x taille (n=1835)



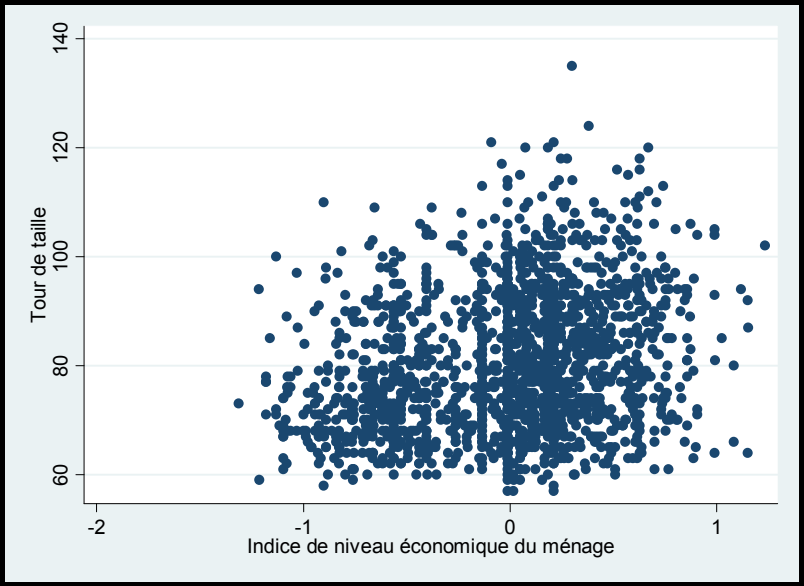
IMC x tour de taille (n=1736)



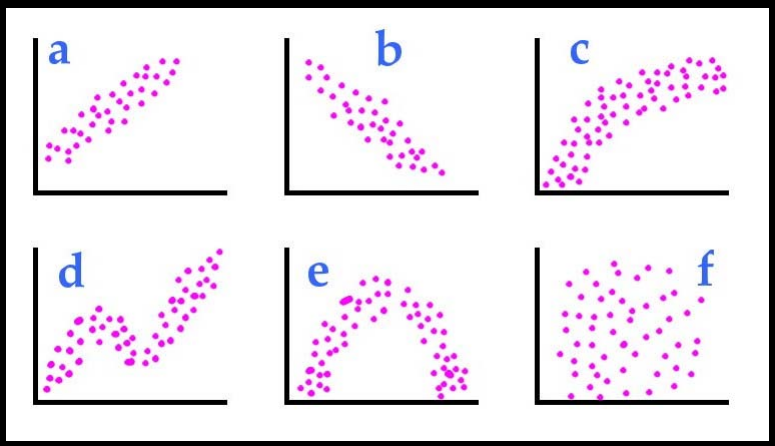
Rapport taille / hanches x age (n=1738)



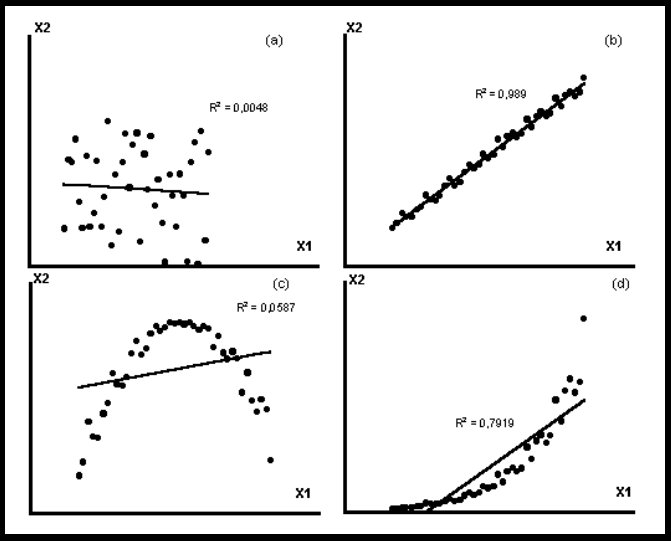
Rapport tour de taille x niveau économique ménage (n=1694)



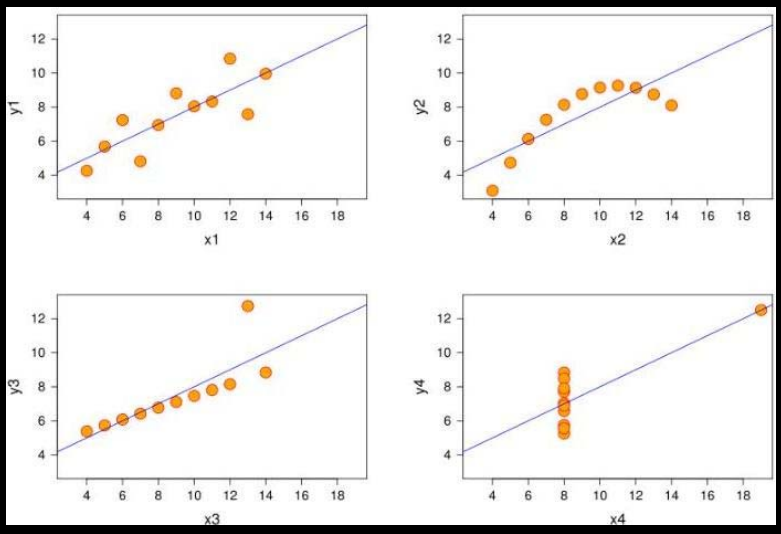
Différentes formes d'association



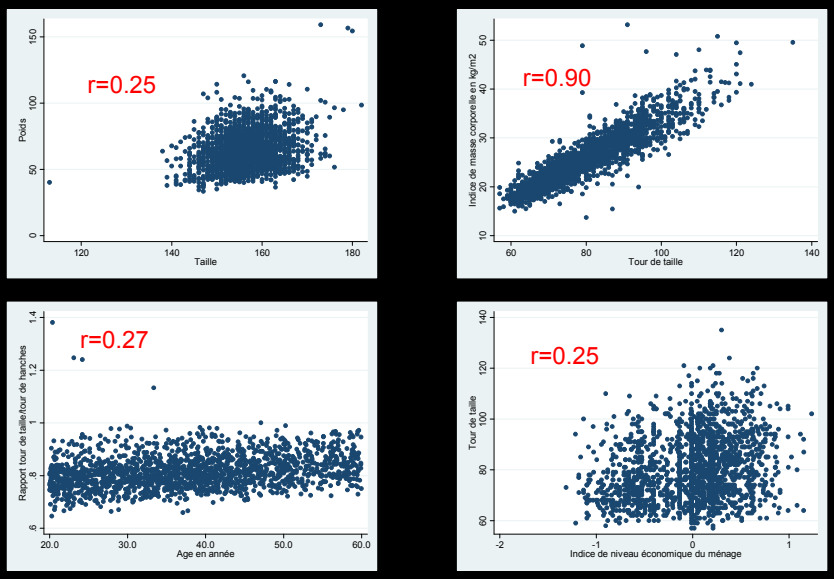
Coefficient de corrélation **linéaire !**



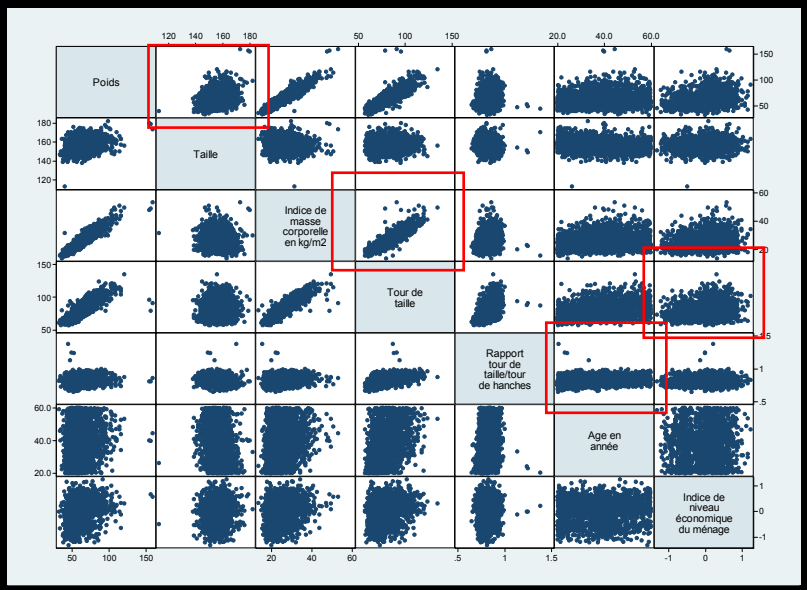
Coefficient de corrélation linéaire ! (r= 0.816)



Coefficients de corrélation linéaire



Matrice des graphiques des variables deux à deux



Matrice des corrélations linéaires des variables deux à deux
(n=1692: données complètes)

	poids	taille	imc	tt	rth	age	eco1
poids	1						
taille	0.26	1					
imc	0.93	-0.09	1				
tt	0.85	-0.01	0.90	1			
rth	0.22	-0.12	0.28	0.54	1		
age	0.23	-0.22	0.32	0.42	0.27	1	
eco1	0.30	0.10	0.27	0.25	0.06	0.04	1

Projet Obe Maghreb

Ecole thématique gestion et analyse de données
20 au 29 avril 2010

Gestion et analyse de données d'enquêtes épidémiologiques

Analyse de données

7. Trois variables qualitatives
Facteur de confusion. Ajustement. Effet modificateur.
Exemples utilisés pendant le cours

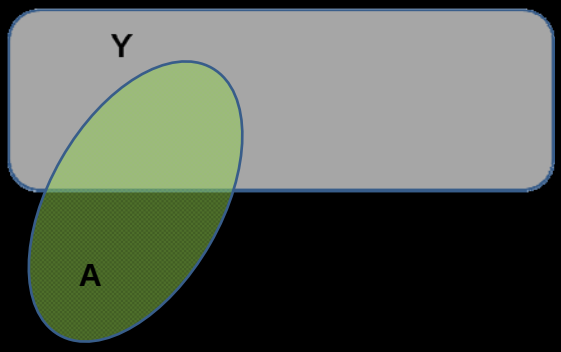
Institut de recherche pour le développement

Pierre Traissac

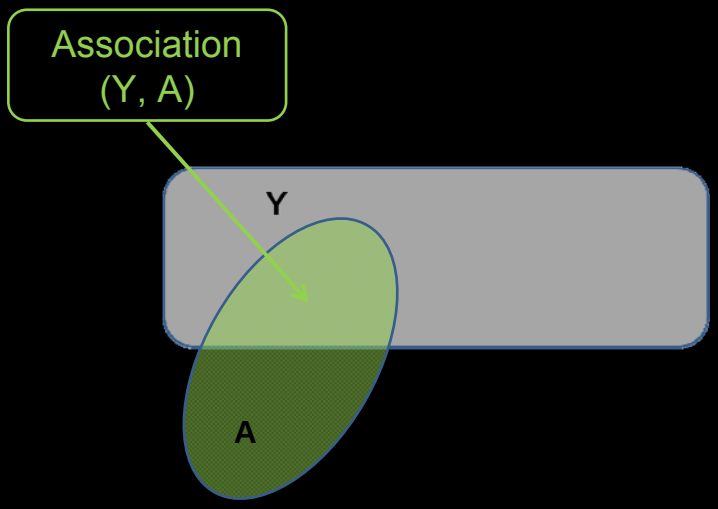
UMR 204 « Prévention des malnutritions et pathologies associées »
IRD, Montpellier, France

Variables explicatives indépendantes

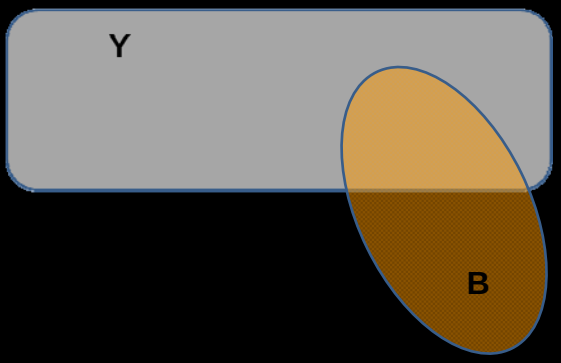
Méthodologie – Ajustement (1)
Variables explicatives indépendantes



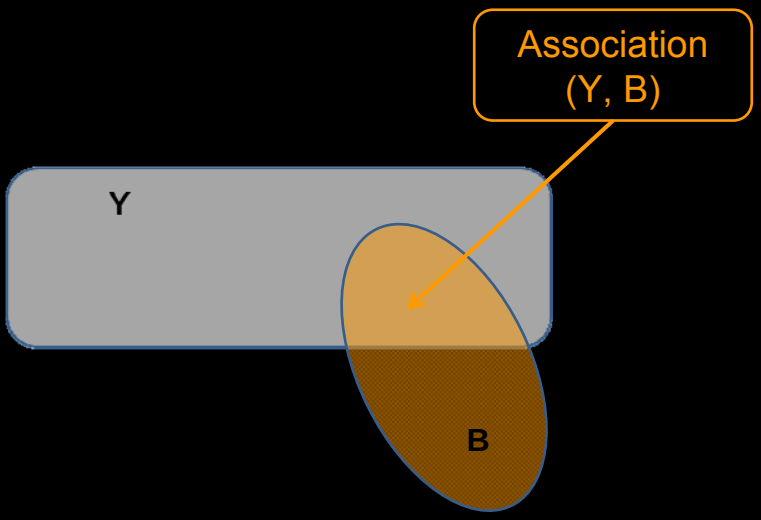
Ajustement
Variables explicatives indépendantes



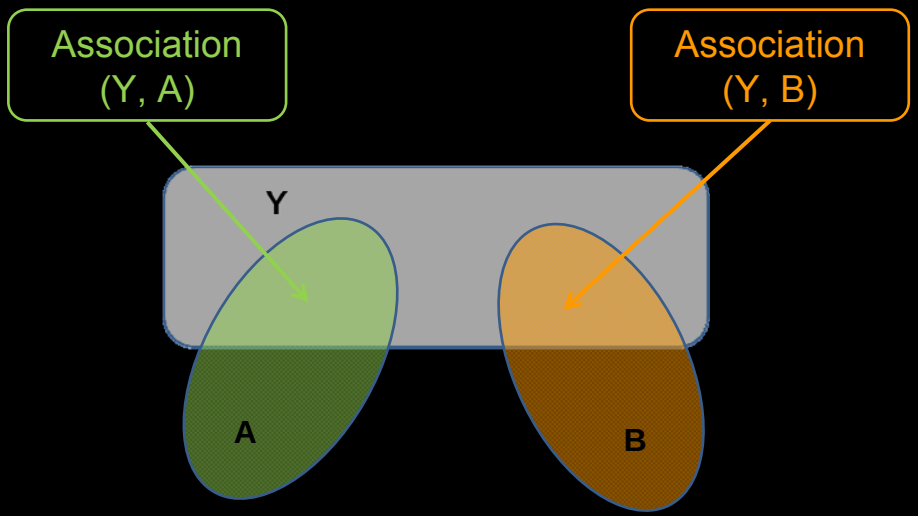
Méthodologie – Ajustement (1)
Variables explicatives indépendantes



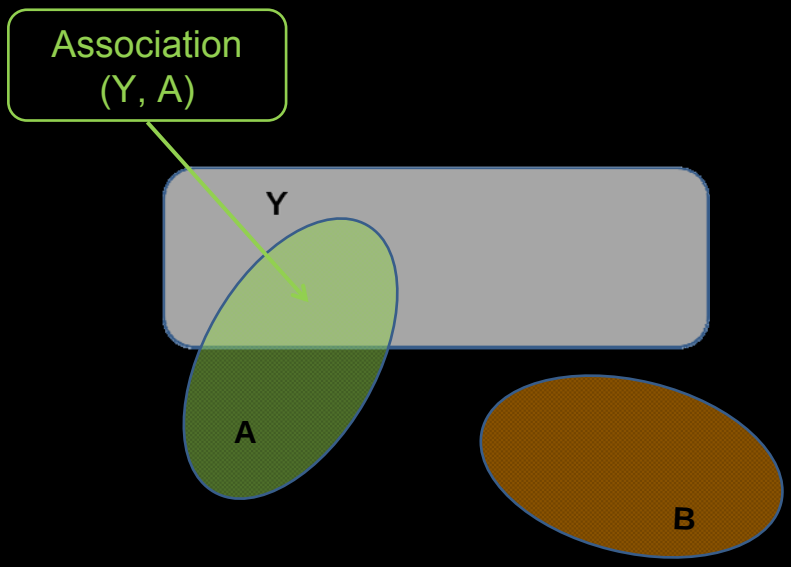
Ajustement
Variables explicatives indépendantes



Ajustement
Variables explicatives indépendantes

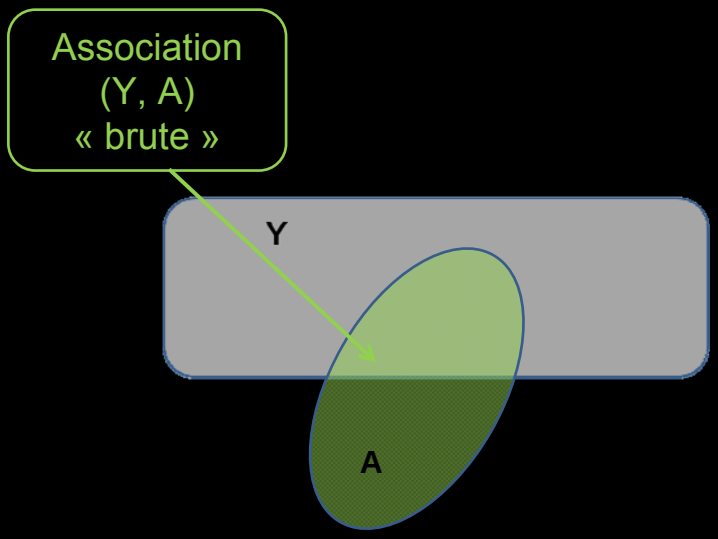


Ajustement
Variables explicatives indépendantes

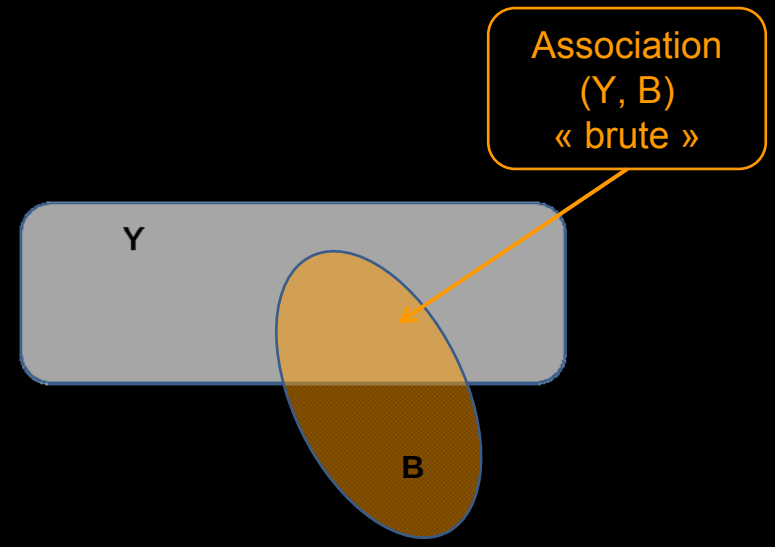


Ajustement
Variables explicatives NON indépendantes

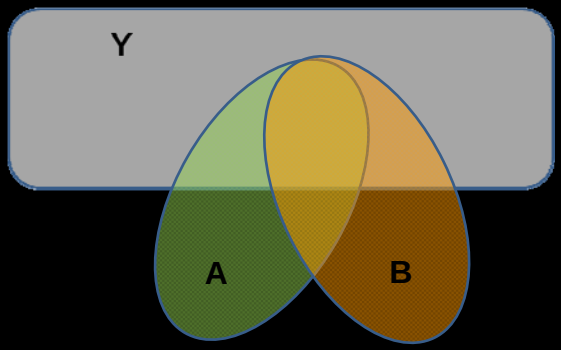
Ajustement
Variables explicatives NON indépendantes



Ajustement
Variables explicatives NON indépendantes

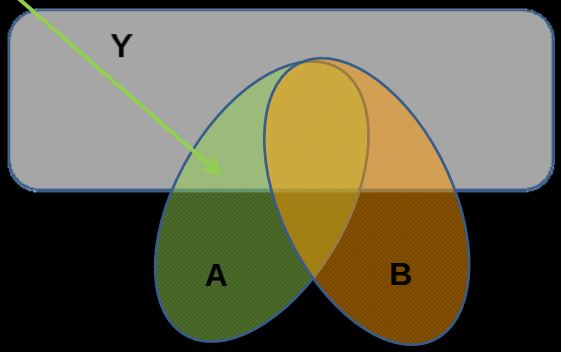


Méthodologie – Ajustement (2)
Variables explicatives NON indépendantes



Méthodologie – Ajustement (2)
Variables explicatives NON indépendantes

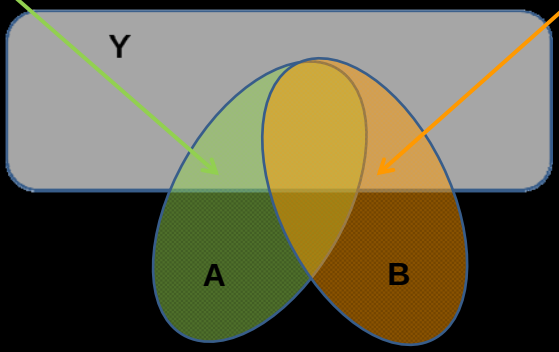
Association
(Y, A)/B
« Ajustée pour B »



Ajustement
Variables explicatives NON indépendantes

Association
(Y, A)/B
« Ajustée pour B »

Association
(Y, B)/A
« Ajustée pour A »



Ajustement