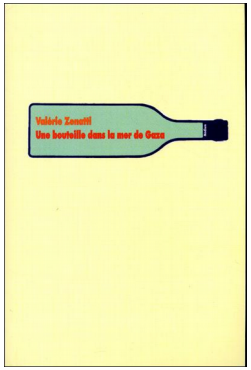
Introduction à la statistique

Statistique exploratoire et statistique inférentielle



*Mais les probabilités, les statistiques, c'est bon pour les maths, **la biologie**, ce sont des chiffres sur du papier.*

Source : *Une bouteille dans la mer de Gaza*, Delphine Zenatti

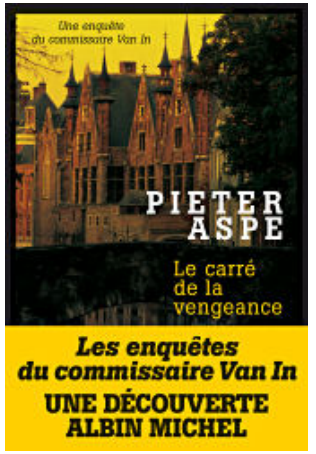
INSTITUT de MATHEMATIQUES de TOULOUSE

Sébastien Déjean

GENOTOUL Biostat

math.univ-toulouse.fr/~sdejean

*Selon le profil que j'ai établi en me basant sur les deux fax, le plus vieux des ravisseurs est très certainement **universitaire**. Il doit être **ingénieur** ou **mathématicien**. Je ne serais pas étonné si on me disait que c'est un **statisticien** ou un spécialiste du **calcul des probabilités**.*



Source : Le carré de la vengeance, Pieter Aspe

Statistique exploratoire

1) Indicateurs statistiques

2) Représentations graphiques

Sébastien Déjean - math.univ-toulouse.fr/~sdejean

Indicateurs statistiques

- Indicateurs pour 1 série de données
  - Position / Tendance centrale
  - Dispersion
- Indicateurs pour 2 séries de données

Objectif : résumer une série de chiffres par un ou plusieurs chiffres (mais pas beaucoup)

Indicateurs de tendance centrale

Moyenne : somme des observations divisées par le nombre d'observations $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Médiane : Valeur qui sépare l'échantillon en 2 sous-ensembles de tailles égales.

Mode : Valeur la plus fréquente dans un échantillon.

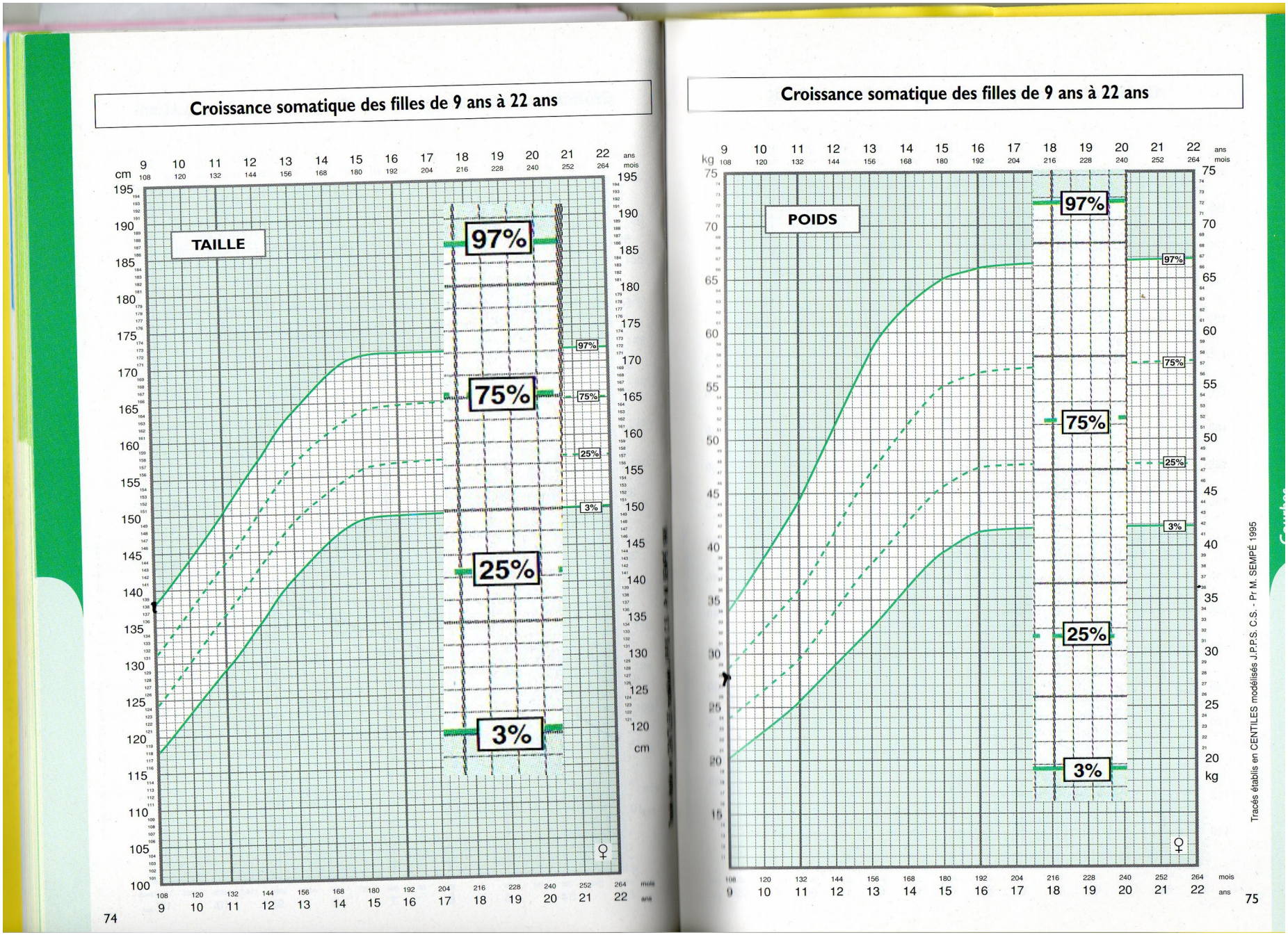
Quartiles : **3** valeurs qui sépare l'échantillon en **4** sous-ensembles de tailles égales.

Déciles : **9** valeurs qui sépare l'échantillon en **10** sous-ensembles...

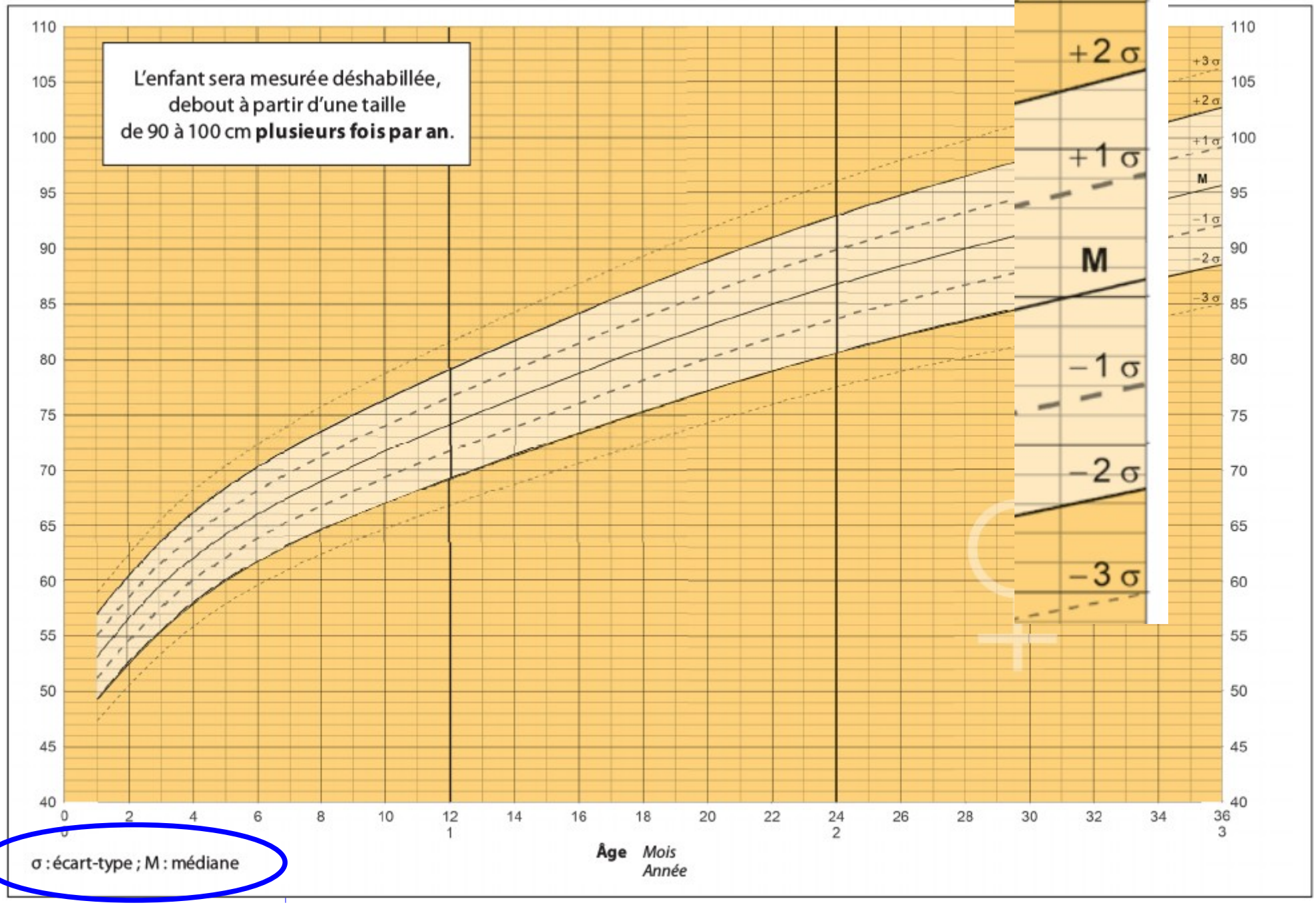
Percentiles : **99** valeurs qui sépare l'échantillon en **100** sous-ensembles...

Quantiles : généralise les précédents

Percentiles ?



Carnet de santé



Moyennes

Moyenne arithmétique : somme des observations divisées par le nombre d'observations
= *moyenne au sens de la somme comme dans une suite... arithmétique* $u_{n+1} = u_n + r$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Moyenne géométrique : racine $n^{ème}$ du produit des observations
= *moyenne au sens du produit comme dans une suite... géométrique* $u_{n+1} = r.u_n$

$$\bar{x}^G = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Exemple : nombre de participants à une manifestation

Les manifestants étaient 10000 selon les organisateurs et 100 selon les autorités. Supposons que chaque camp a volontairement multiplié ou divisé le chiffre réel (=1000) par 10.

La moyenne arithmétique vaut (10000+100)/2 = 5050
La moyenne géométrique vaut $(10000*100)^{1/2}$ = 1000 (on retrouve bien la valeur réelle)

Et aussi : moyenne harmonique, moyenne quadratique...

Moyenne et médiane

Exemple : supposons que l'entreprise compte 12 personnes :

- 8 ouvriers : ~ 1000 € / personne
- Un chef d'atelier : 2000 €
- Un directeur technique : 4000 €
- Un directeur des ressources humaines : 6000 €
- Un directeur général : 10 000 €

```
  970
  975
  980
  985
  990
  995
 1005
 1100
 2000
 4000
 6000
10000

Moyenne : 2500
Médiane : 1000
```

Salaire moyen : **2500 €** - Salaire médian : **1000 €**

Comment augmenter le salaire moyen dans cette société ?

Solution 1 : augmenter tous les salaires de x %

Solution 2 : le dirigeant augmente son salaire

Solution 3 : « remercier » quelques postes de bas salaire

Solution 1

```
1018.50
1023.75
   1029
1034.25
1039.50
1044.75
1055.25
   1155
   2100
   4200
   6300
  10500

Moyenne : 2625
Médiane : 1050
```

Solution 2

```
 970
 975
 980
 985
 990
 995
1005
1100
2000
4000
6000
12000

Moyenne : 2666
Médiane : 1000
```

Solution 3

```
 980
 985
 990
 995
1005
1100
2000
4000
6000
12000

Moyenne : 2805.50
Médiane : 1052.50
```

Revue de presse

latribune.fr | 14/03/2013

L'an dernier, **le salaire moyen brut a progressé de 2,1%**, soit près d'un point de plus que l'inflation (+1,2%). Mais cette hausse ralentit.

Petite éclaircie pour le pouvoir d'achat des Français. L'Agence centrale des organismes de Sécurité sociale (Acoss) a annoncé les chiffres sur l'évolution du **salaire moyen** brut en France. En 2012, il a augmenté de 2,1% pour s'établir à 2.410 euros brut, soit 0,9 point de plus que l'inflation qui s'est tenu à 1,2%. Mais il apparaît que derrière le chiffre de cette progression, se cache en réalité, un ralentissement tendanciel. Au premier trimestre, la hausse était de 0,6%, puis de 0,5% aux deux trimestres suivants, pour terminer sur une progression de 0,4% au dernier trimestre.

Les effectifs salariés en légère baisse

La progression du salaire moyen s'explique en grande partie par la hausse de la masse salariale (+1,7%) alors que le nombre de salariés du secteur privé s'est légèrement contracté (-0,5%, soit 94.000 pertes nettes d'emploi). L'Acoss note que la hausse de la masse salariale s'inscrit en net ralentissement puisqu'elle avait été de 3,5% l'année précédente. La baisse des effectifs est davantage marquée dans le secteur immobilier (-1,1%). En revanche, l'organisme note une hausse des effectifs dans l'action sociale (+0,4%), et dans la restauration (+0,2%). En termes géographiques, "les régions du nord-est" enregistrent les plus fortes baisses d'effectifs salariés. L'emploi a toutefois progressé dans les Dom, en Corse, en Midi-Pyrénées et en Ile-de-France.

Humour...

Durant les élections de 1956, un électeur apostropha le candidat Adlai E. Stevenson :

« Senateur, vous avez pour vous toutes les personnes intelligentes ». Stevenson répondit : « Ce n'est pas suffisant, nous avons besoin d'une majorité !»

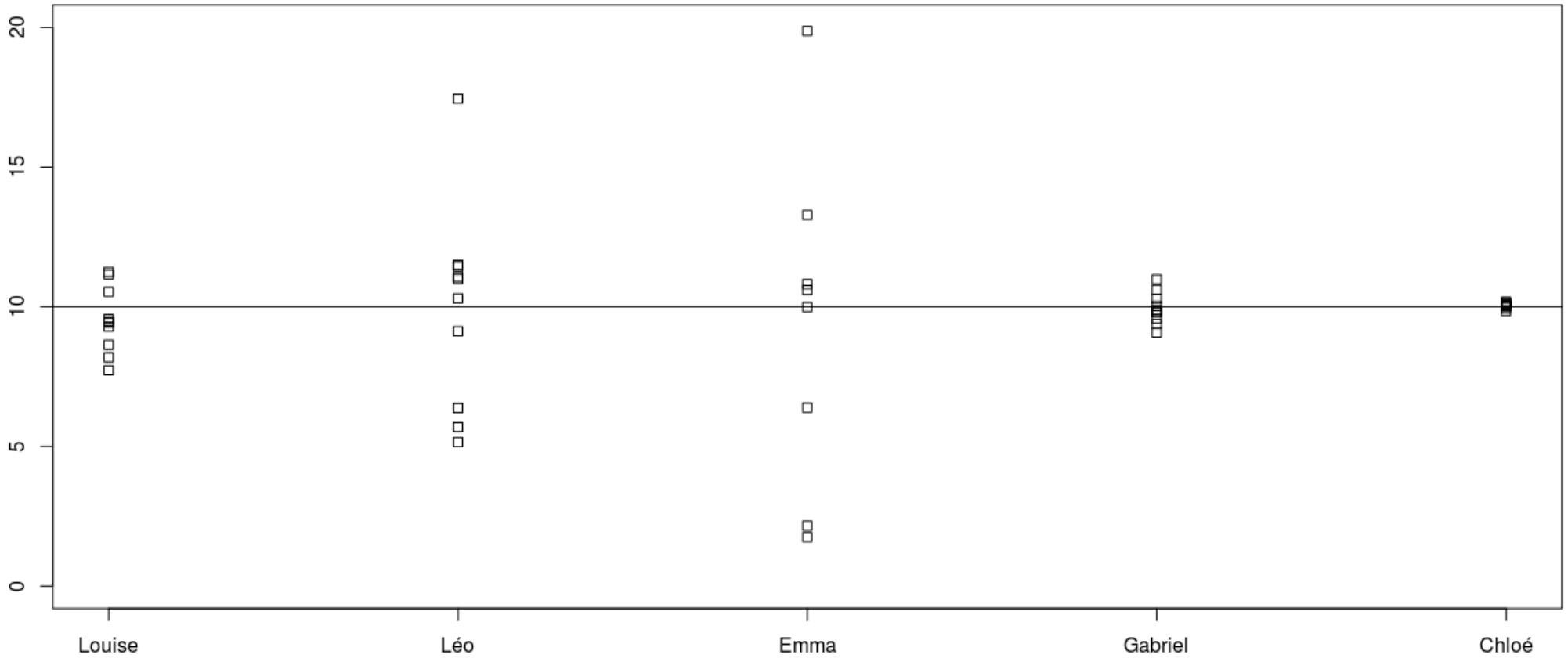
« La majorité des français sont plus cons que la moyenne »

Gustave Parking

Que se passe-t-il lorsqu'une blonde française passe la frontière pour aller en Belgique ?

Le QI moyen des 2 pays augmente !

Limites des indicateurs de position



Ce graphique représente les notes obtenues par 5 élèves dans 10 matières différentes. Les 5 élèves ont une moyenne (et une note médiane) à peu près égale à 10. Donneriez-vous la même appréciation (*Élève moyen*) à chacun de ces élèves ?

Indicateurs de dispersion

Variance

$$\mathrm{var}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Moyenne des carrés des écarts à la moyenne

Écart-type

$$\sigma(X) = \sqrt{\mathrm{var}(X)}$$

Racine carrée de la variance

.....

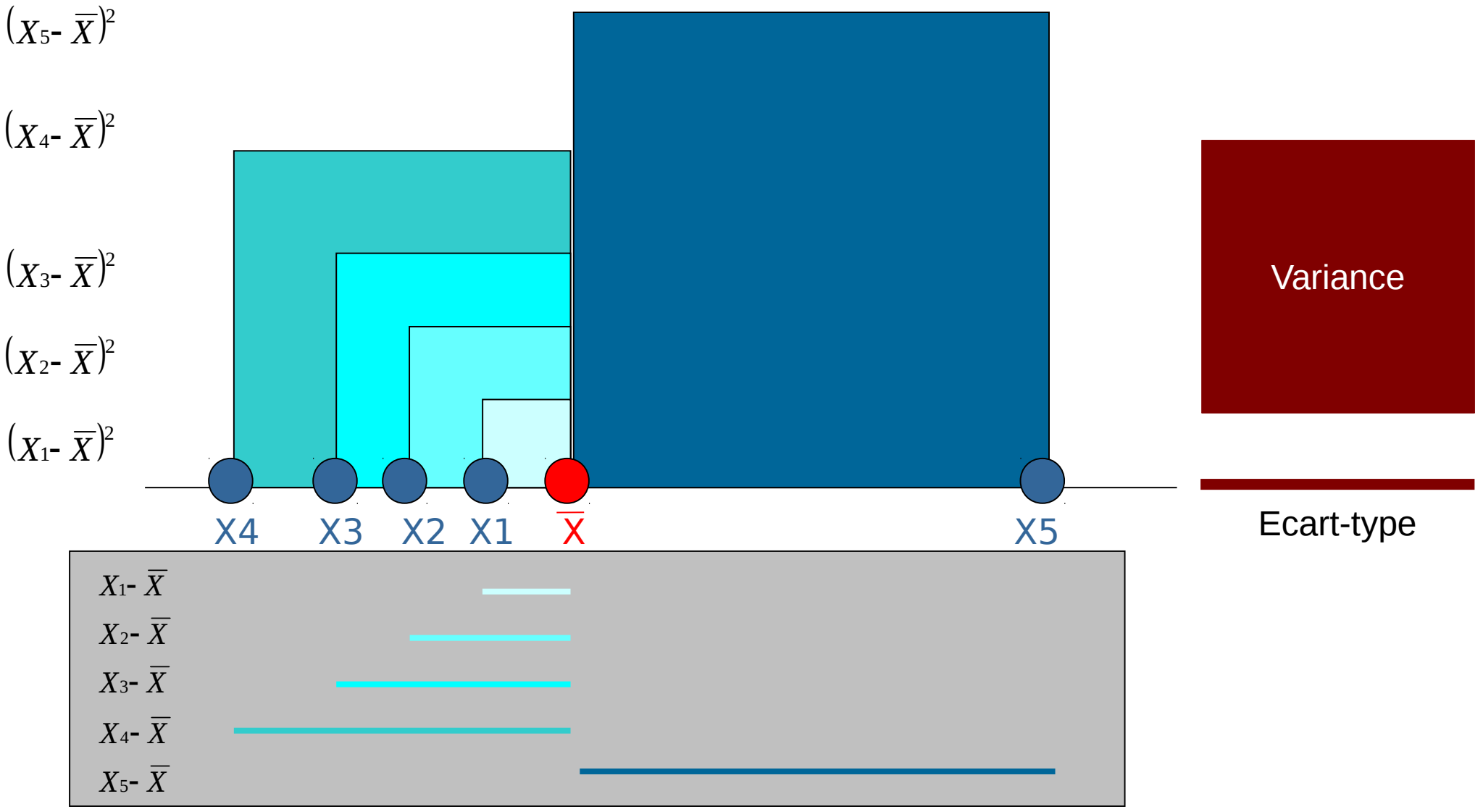
Étendue : différence entre la plus grande et la plus petite valeur d'un échantillon

Espace inter-quartile : différence entre le 1er et le 3ème quartile, correspond à l'étendue de l'échantillon privé de la moitié de ces observations (le ¼ le plus élevé et le ¼ le plus faible)

www.variance.fr

Variance et écart-type

Racine carrée de la moyenne des carrés des écarts à la moyenne



Variance et écart-type

Quelques propriétés de l'écart-type :

- Positif (nul si la série est constante)
- Invariant par translation
- Sensible aux valeurs extrêmes
- **De la même unité que la donnée** (et que la moyenne) :

*Si l'échantillon est constitué de mesures en **m** alors l'écart-type s'exprime également en **m** (tout comme la moyenne) ; ce qui n'est pas le cas de la variance **m²** !*

On peut ainsi additionner moyenne et écart-type (*mais pas moyenne et variance*), ce qui est fondamental pour la construction d'intervalle de confiance.

Revue de presse

Les filles brillent en classe, les garçons aux concours

LE MONDE | 07.09.09 - Article paru dans l'édition du 08.09.09. Philippe Jacqué

Elles obtiennent de meilleurs résultats en cours de scolarité, mais réussissent moins bien les concours des meilleures grandes écoles que les hommes. Raison : les femmes souffriraient plus dans un *"environnement concurrentiel"*.
[...]
Les conclusions de cette étude sont accablantes. Les candidates aux concours de l'école de Jouy-en-Josas (Yvelines) ont beau avoir de meilleurs dossiers que leurs concurrents masculins (mentions au bac supérieures, meilleure représentation dans les bonnes classes préparatoires), elles y réussissent moins bien. Alors que le pourcentage d'hommes et de femmes candidats est équilibré sur les trois années étudiées (50,84 % d'hommes, 49,16 % de femmes), le pourcentage de femmes admissibles tombe à 46,32 %, et celui d'admises à 45,92 %... Pis, après le concours, *"celles qui l'ont réussi obtiennent en première année en moyenne des notes d'examen supérieures à celles de leurs congénères masculins."*
[...]

[…]

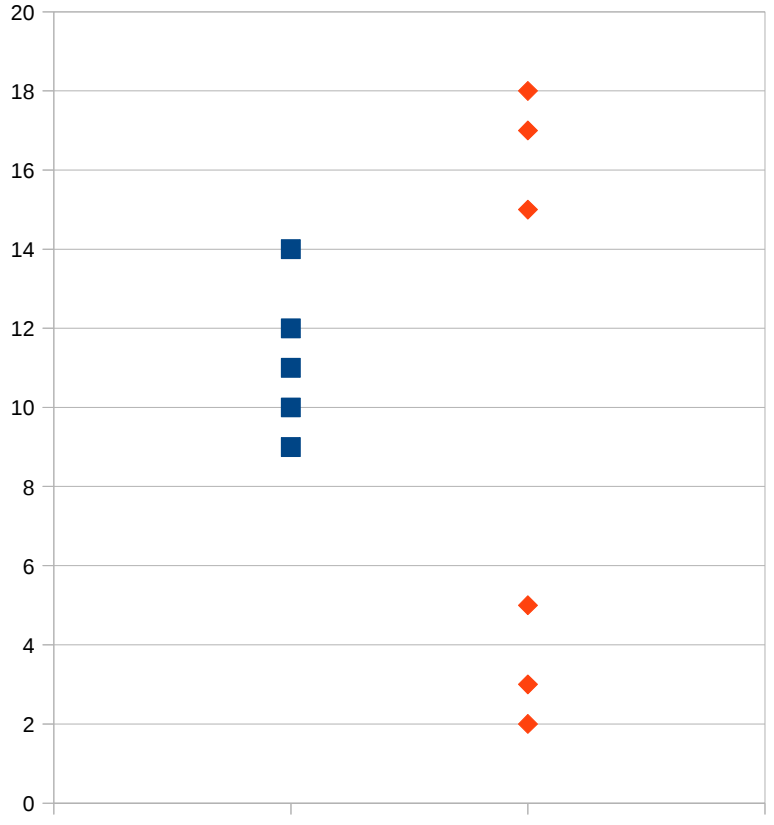
"Autour de la moyenne"

Pour expliquer la moindre réussite des femmes, une rumeur court depuis de nombreuses années : les femmes seraient discriminées aux oraux. *"Si un des jurys d'oral peut avoir des biais, aucune consigne n'est donnée en ce sens"*, assure Frédéric Palomino. *"Nous avons mené cette enquête statistique pour tordre le cou à ce fantasme"*, explique Eloïc Peyrache. *"De plus, quand on regarde les pourcentages de réussite, on voit que c'est à l'écrit que la part des candidates chute le plus."*

Alors comment comprendre ce déséquilibre ? ***"D'un point de vue technique, il semble que la structure du concours HEC crée d'avantage d'hétérogénéité chez les hommes que chez les femmes"*, estime M. Peyrache. Si, *"en moyenne"*, les performances des hommes et des femmes sont similaires, *"les notes des femmes sont concentrées autour de la moyenne, tandis que celles des hommes sont très dispersées avec beaucoup de très bonnes notes et de très mauvaises. Mécaniquement, quand on sélectionne les 380 premiers résultats, on a un peu plus d'hommes"*.**

Illustration

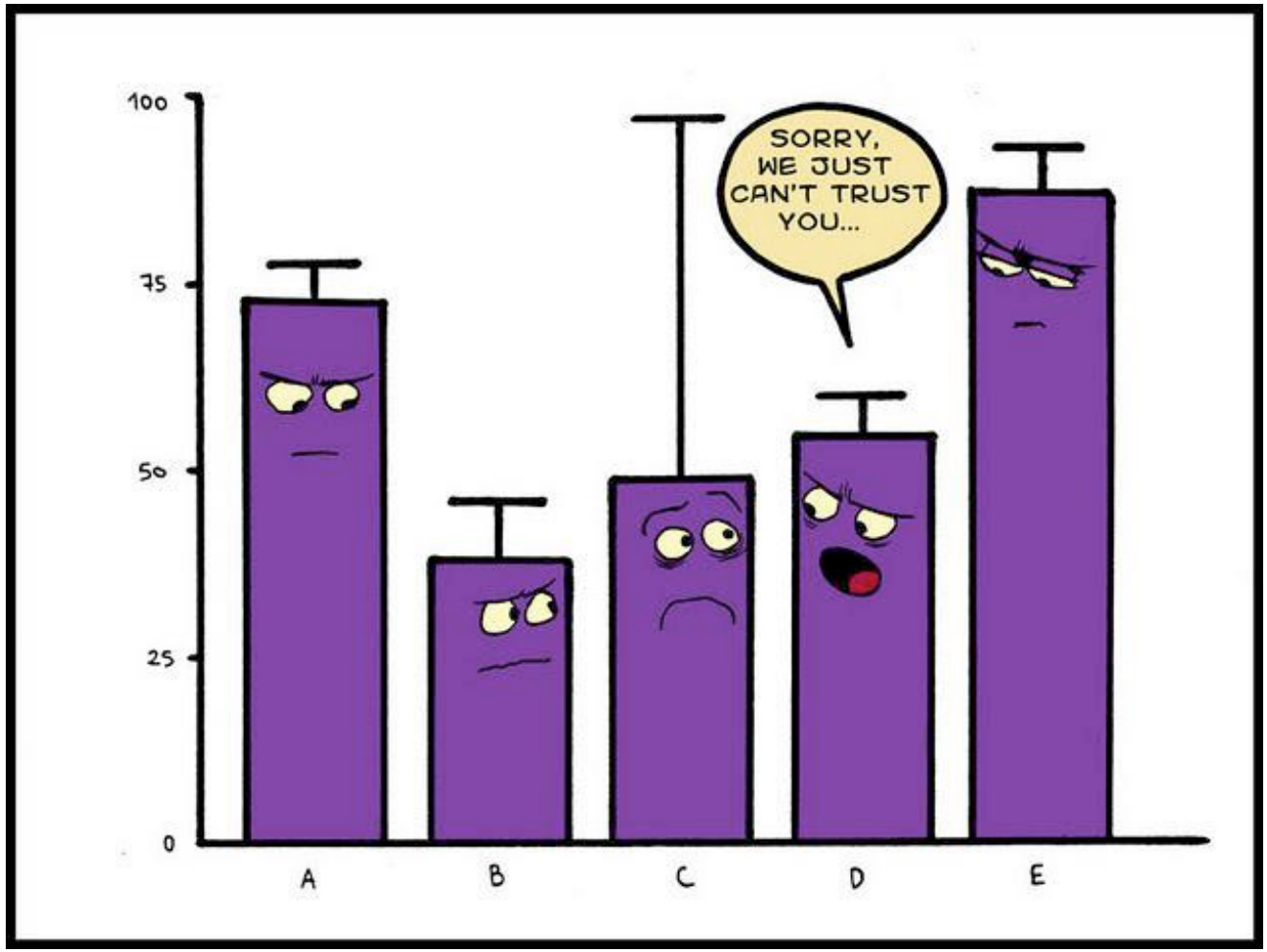
	Femme	Homme
	9	2
	11	18
	10	15
	12	3
	12	17
	14	5
Moyenne	11.33	10
Etendue	5	16
Ecart-type	1.75	7.43
Variance	3.07	55.2



Classement
18
17
15
14
12
12
11
10
9
5
3
2

Humour de statisticien...

Source : xkcd.com



LES STATISTIQUES
C'EST PAS AUTOMATIQUE

Parlez-en à un-e statisticien-ne

However, if n is very small (for example n = 3), rather than showing error bars and statistics, it is better to simply plot the individual data points.

Cumming, G., Fidler, F., & Vaux, D. L. (2007).
Error bars in experimental biology.
The Journal of Cell Biology, 177(1), 7–11.

Centrer / réduire

- **Centrer** : retrancher la moyenne

- **Réduire**(*) : diviser par l'écart-type

  → réduire revient parfois à dilater les données, si l'écart-type est inférieur à 1...

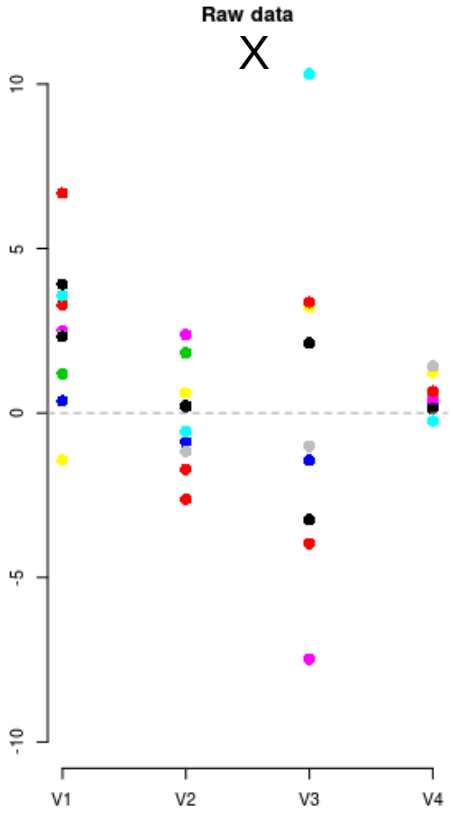
- Permet d'exprimer des variables différentes sur une échelle commune, en les débarrassant de leurs unités physiques : les observations s'expriment en **nombres d'écart-type par rapport à la moyenne.**

- Après centrage-réduction, la moyenne des observations est nulle et l'écart-type vaut 1 (ainsi que la variance).

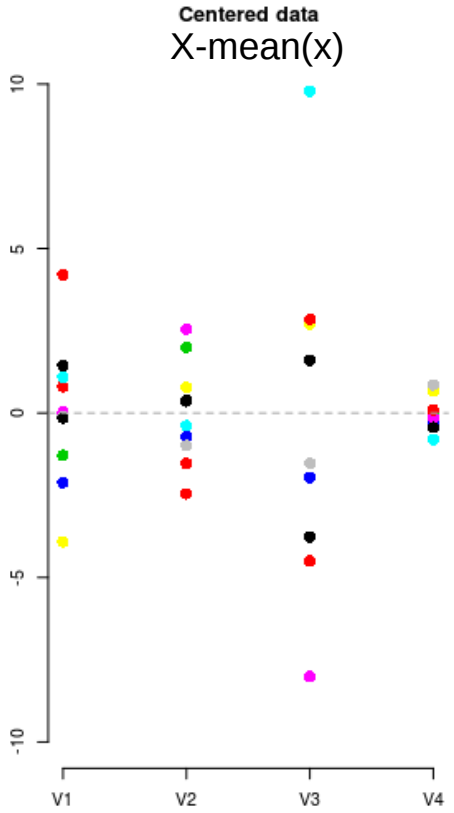
$$Z_i = \frac{X_i - \overline{X}}{\sigma_X}$$

- Appelé parfois « z-transformation » ou « z-score »

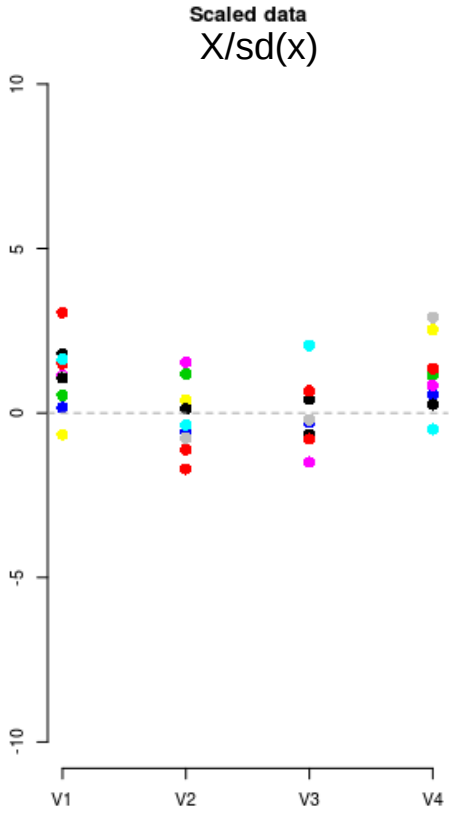
Centrer / réduire



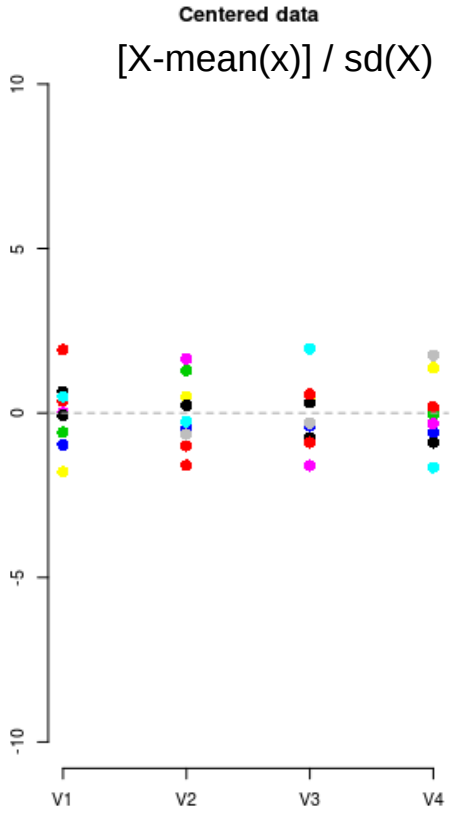
	v1	v2	v3	v4
1	3.9	0.2	-3.2	0.6
2	3.3	-1.7	-4.0	0.6
3	1.2	1.8	3.3	0.6
4	0.4	-0.9	-1.4	0.3
5	3.6	-0.6	10.3	-0.2
6	2.5	2.4	-7.5	0.4
7	-1.4	0.6	3.2	1.2
8	2.4	-1.2	-1.0	1.4
9	2.3	0.2	2.1	0.1
10	6.7	-2.6	3.4	0.7
Mean	2.5	-0.2	0.5	0.6
S.D.	2.2	1.5	5.0	0.5



	v1	v2	v3	v4
1	1.4	0.4	-3.8	0.1
2	0.8	-1.5	-4.5	0.0
3	-1.3	2.0	2.8	0.0
4	-2.1	-0.7	-2.0	-0.3
5	1.1	-0.4	9.8	-0.8
6	0.0	2.5	-8.0	-0.2
7	-3.9	0.8	2.7	0.7
8	-0.1	-1.0	-1.5	0.9
9	-0.2	0.4	1.6	-0.4
10	4.2	-2.5	2.8	0.1
Mean	0	0	0	0
S.D.	2.2	1.5	5.0	0.5



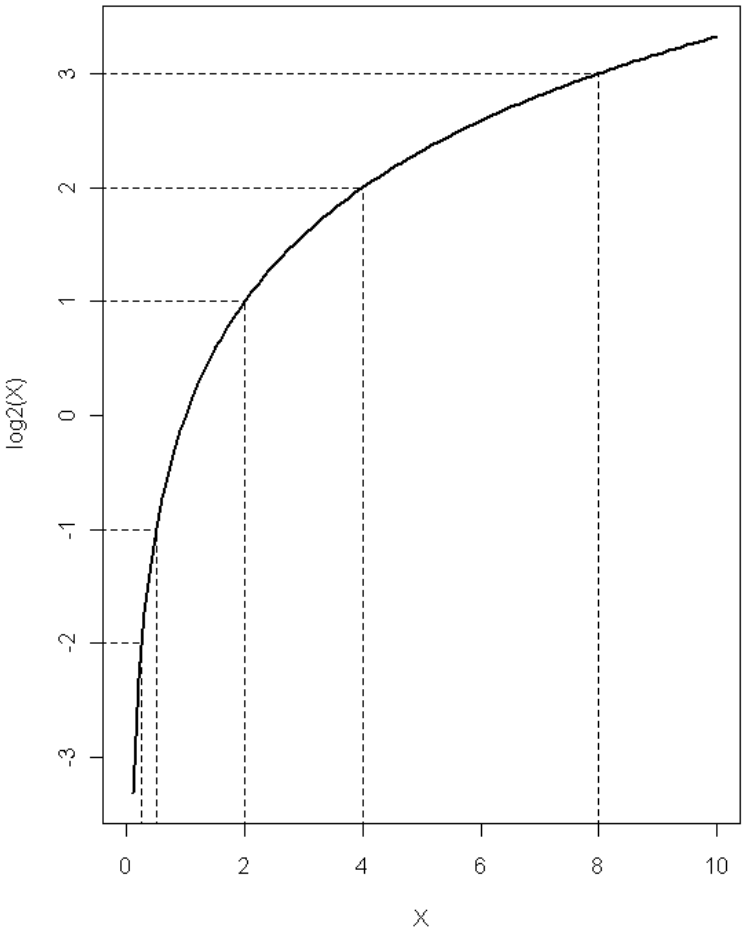
	v1	v2	v3	v4
1	1.1	0.1	-0.6	0.8
2	1.0	-1.1	-0.8	0.8
3	0.3	1.2	0.7	0.7
4	0.1	-0.6	-0.3	0.4
5	1.0	-0.4	2.0	-0.3
6	0.7	1.5	-1.5	0.5
7	-0.4	0.4	0.6	1.6
8	0.7	-0.7	-0.2	1.8
9	0.7	0.1	0.4	0.2
10	2.0	-1.7	0.7	0.9
Mean	1.1	-0.1	0.1	1.2
S.D.	1	1	1	1



	v1	v2	v3	v4
1	0.7	0.3	-0.8	0.1
2	0.4	-1.0	-0.9	0.0
3	-0.6	1.3	0.6	0.0
4	-1.0	-0.5	-0.4	-0.6
5	0.5	-0.3	2.0	-1.7
6	0.0	1.7	-1.6	-0.3
7	-1.8	0.5	0.5	1.4
8	-0.1	-0.6	-0.3	1.7
9	-0.1	0.2	0.3	-0.9
10	1.9	-1.6	0.6	0.2
Mean	0	0	0	0
S.D.	1	1	1	1

Conversion « log »

X	0,125 $= 2^{-3}$	0,25 $= 2^{-2}$	0,5 $= 2^{-1}$	1 $= 2^{0}$	2 $= 2^{1}$	4 $= 2^{2}$	8 $= 2^{3}$
$\log_2(X)$	-3	-2	-1	0	1	2	3



- Utile pour la conversion de ratio
- Exemple : Une sur-expression double et une sous-expression de moitié sont rendues symétriques par rapport à 0
- Pour les p-values, on aura plus intérêt à utiliser $\log_{10}$.
- La fonction réciproque de « log » est la fonction puissance :

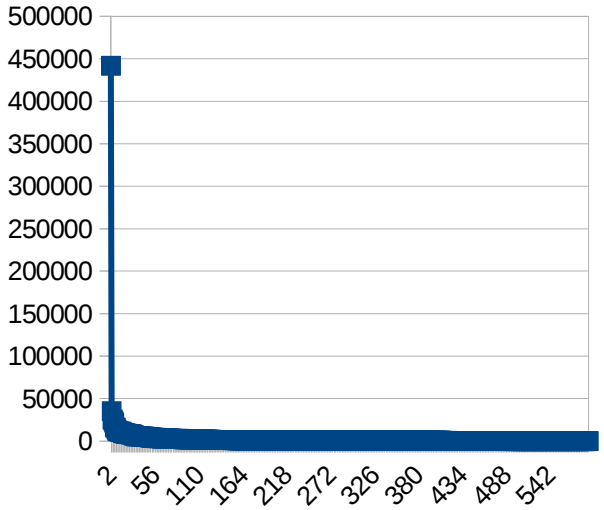
$$Y = \log_2(X) \leftrightarrow X = 2^Y$$

$$Y = \log_{10}(X) \leftrightarrow X = 10^Y$$

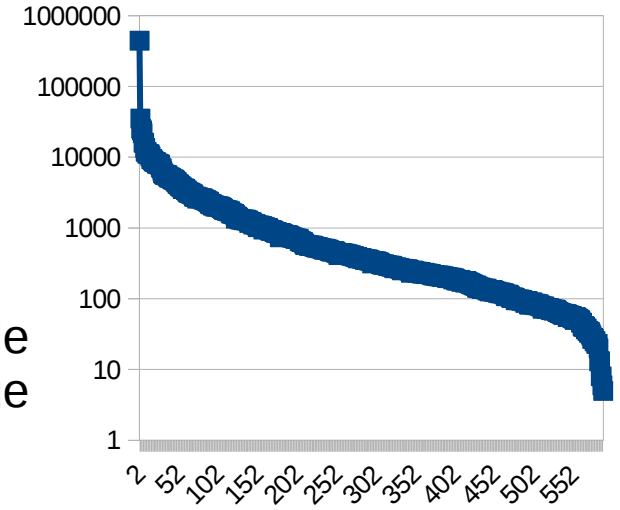
$$Y = \ln(X) \leftrightarrow X = e^Y = \exp(Y)$$

Conversion « log »

Populations légales des communes en vigueur au 1er janvier 2013
Mise à jour : décembre 2012
en habitant
Champ : Département de la Haute-Garonne, limites territoriales en vigueur au 1er janvier 2012
Date de référence statistique : 1er janvier 2010
Source : Insee, Recensement de la population 2010



Échelle linéaire



Échelle logarithmique

Commune	Population	Ordre de grandeur || Log10
Toulouse	441 802	5.65
Colomiers	35 186	4.55
Tournefeuille	25 340	4.40
Muret	23 864	4.38
...		
Castanet-Tolosan	11 033	4.04
Saint-Orens...	10 918	4.04
Saint-Jean	10 259	4.01
Revel	9 361	3.97
Portet-sur-Garonne	9 435	3.97
Auterive	9 107	3.96
...		
La Magdelaine-sur-T/	1 006	3.00
Grépiac	990	2.99
Landorthe	946	2.98
Vigoulet-Auzil	944	2.97
...		
Belbèze-de-Lauragais	104	2.02
Saint-Germier	103	2.01
Seyre	102	2.01
Gouzens	95	1.98
Lourde	98	1.99
Pouze	97	1.99
...		
Saccourvielle	13	1.11
Cirès	13	1.11
Bourg-d'Oueil	8	0.90
Trébons-de-Luchon	8	0.90
Caubous	6	0.78
Baren	5	0.70

Ratio, différence, taux

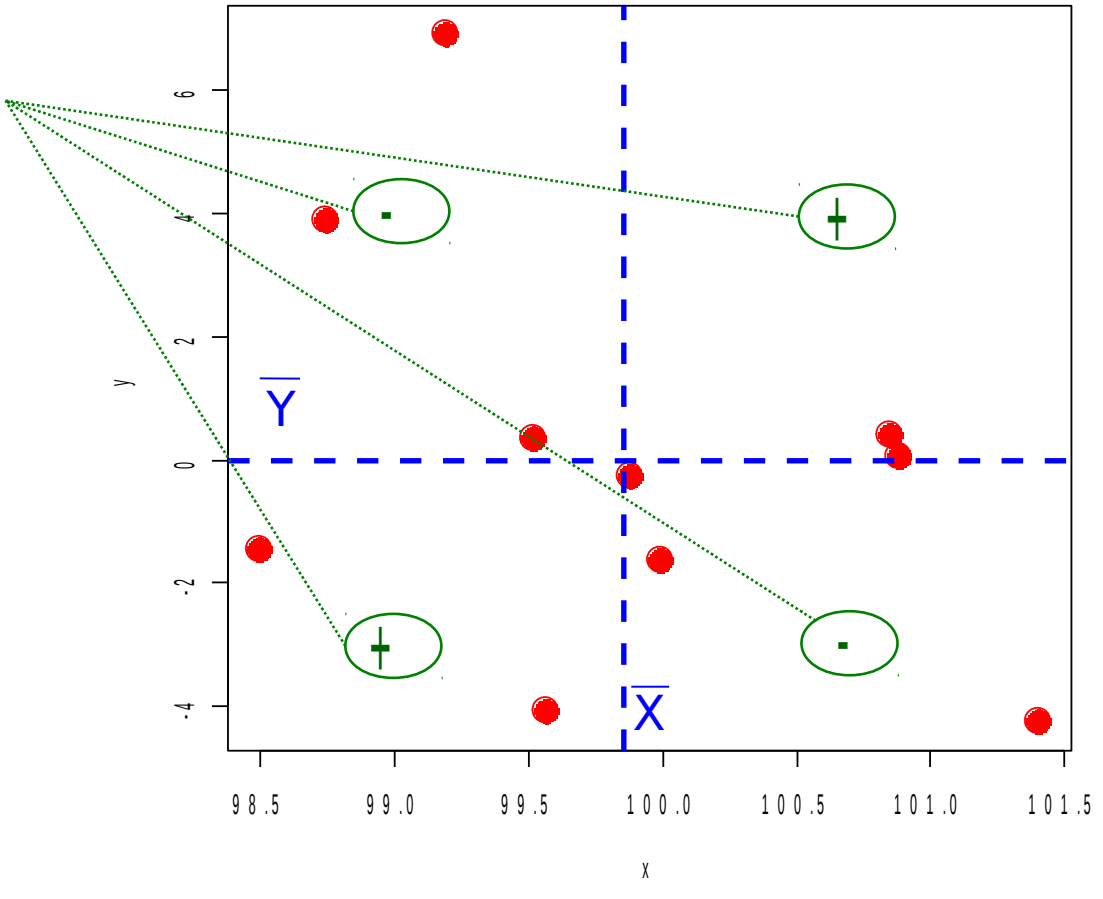
	Indiv1	Indiv2	Indiv3	Indiv4	Indiv5
Mesure temps 1	10	100	100	1000	50
Mesure temps 2	20	200	110	1100	200
Différence (T2-T1)	10	100	10	100	150
Ratio (T2/T1)	2	2	1.1	1.1	4
Taux (T2-T1)/T1	1 100 %	1 100 %	0.1 10 %	0.1 10 %	3 300 %

Indicateurs statistiques 2D

Covariance $$\mathrm{cov}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$$

cov(X,X)=var(X)

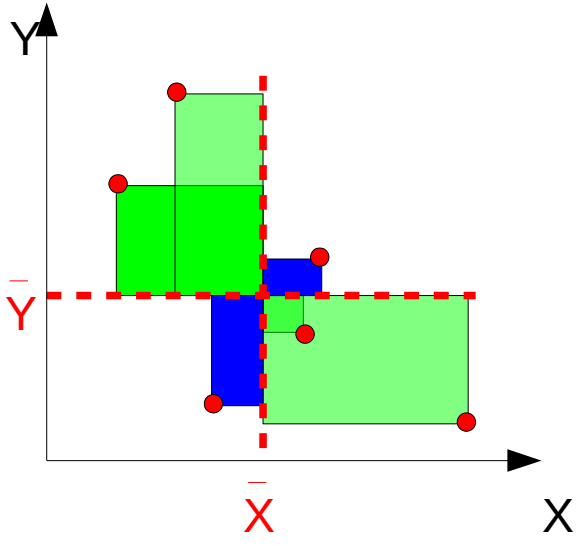
Signe du produit (Xi-$\bar{X}$)(Yi-$\bar{Y}$)



Intuitivement :
- Si les **+** l'emportent
→ liaison linéaire positive

- Si les **–** l'emportent
→ liaison linéaire négative

*Sur cet exemple : cov(X,Y)=**-1.36***



La covariance dépend des unités de mesure ➔ coefficient de corrélation

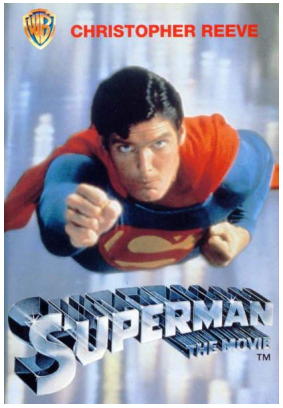
Corrélation

- Coefficient de corrélation **linéaire** de Pearson

$$\rho(X,Y) = cov(X,Y)/(\sigma_X \sigma_Y)$$

- Coefficient de corrélation de Spearman
  ➔ *Robustesse due au travail sur les rangs*

$$\rho_s(X,Y) = \rho(RX,RY) = 1 - 6\frac{\sum_{i=1}^{n} d^2}{n(n^2-1)}$$



(1) *calcul des rangs*

(2) *différence des rangs*

(3) *carrés des différences des rangs*

(4) *somme des carré des différences des rangs*

			(1)		**(2)**	**(3)**
X	Y		RX	RY	d = RX-RY	d²
20,6	20,7		6	7	-1	1
21,6	21,8		2	2	0	0
18,8	20,4		9	8	1	1
20,8	21,1		3	4	-1	1
17,5	18,3		10	10	0	0
19,5	18,9		7	9	-2	4
20,8	21,1		4	4	0	0
20,6	21,2		5	3	2	4
19,2	20,9		8	6	2	4
22,2	22,9		1	1	0	0
					Somme	**15**

(4)

Coefficient(s) de corrélation

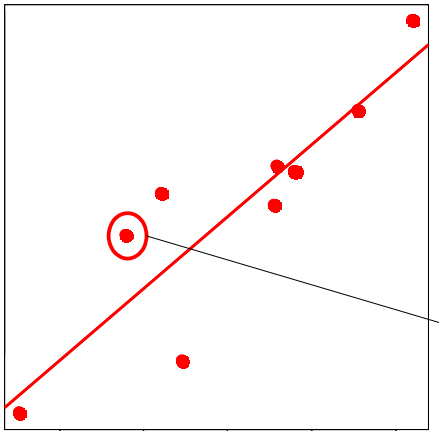
Quelques propriétés des coefficients de corrélation :

- Coefficient de corrélation de Pearson : relation **linéaire**
- Coefficient de corrélation de Spearman (rangs) : relation **monotone**
- Compris entre –1 et 1.
- Les valeurs extrêmes –1 et 1 indique des corrélations parfaites entre les 2 variables.
- Si le coefficient est **positif** : quand une variable est élevée, l'autre l'est également. Quand une variable est faible, l'autre l'est également.
- Si le coefficient est **négatif** : quand une variable est élevée (resp. faible), l'autre est faible (resp. élevée).

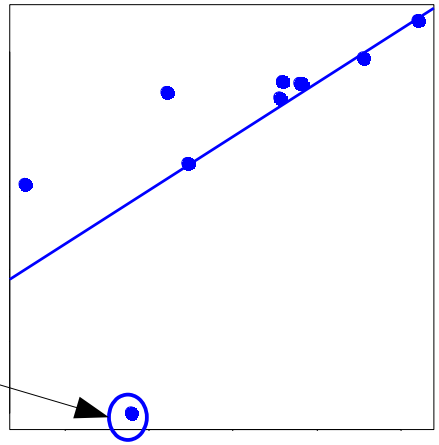
Un exemple de corrélation négative
Ma vie et celle du commissaire Flores avaient suivi des lignes à la fois divergentes et concomitantes : il montait et je descendais dans une corrélation non fortuite, attendu que ses mérites se fondaient généralement sur mes échecs.
Eduardo Mendoza (traduction de Delphine Valentin), Les égarements de mademoiselle Baxter

Corrélation : exemples

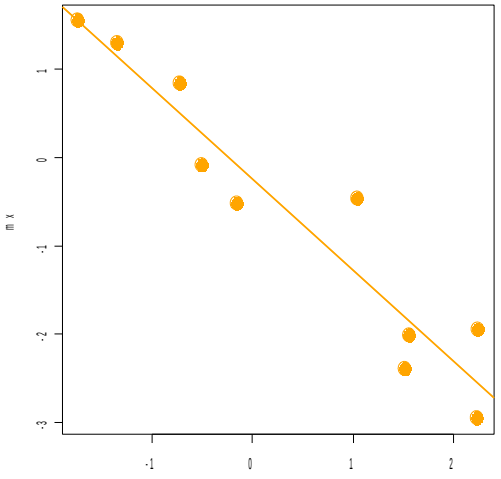
ρ : Pearson - ρ_s : Spearman



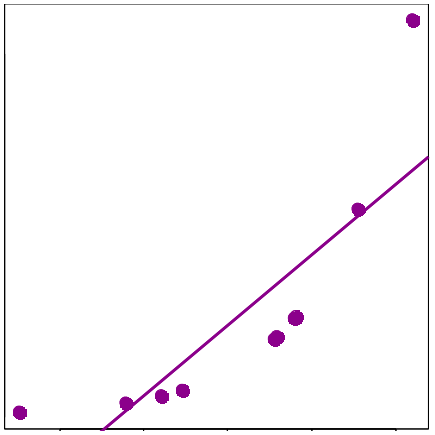
ρ=0.884 - ρ_s=0.9



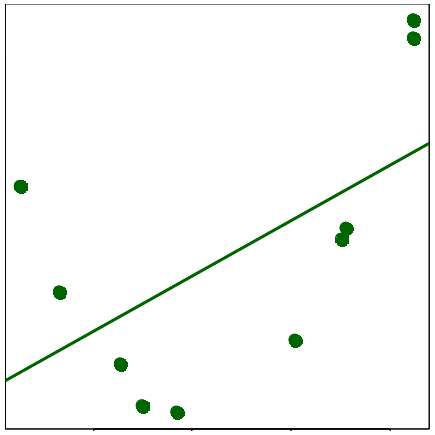
ρ=0.676 - ρ_s=0.912



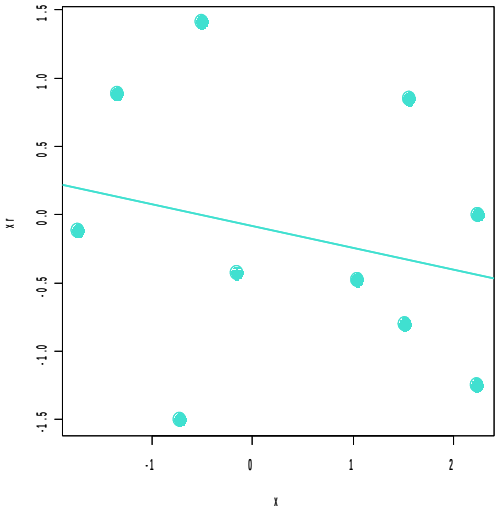
ρ=-0.954 - ρ_s=-0.903



ρ = 0.822 - ρ_s=1



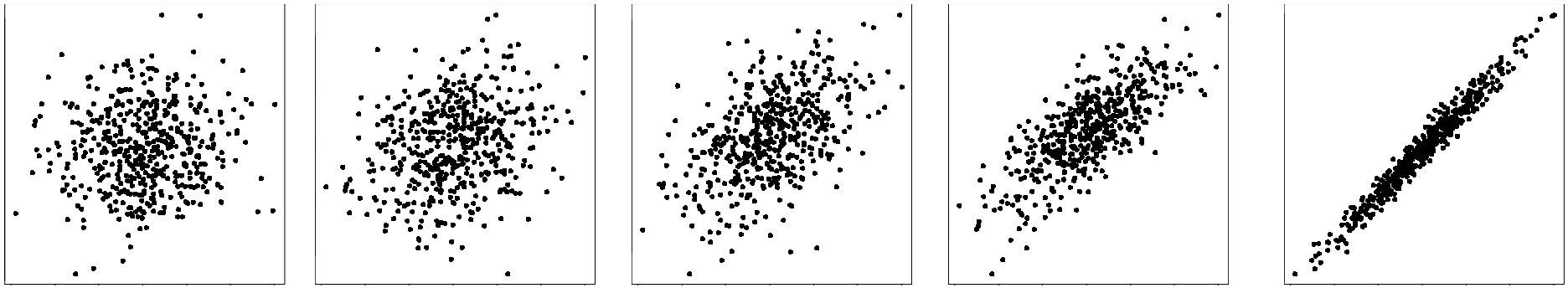
ρ=0.584 - ρ_s=0.491



ρ=-0.248 - ρ_s=-0.164

Corrélation : exemples

0.05 0.26 0.49 0.75 0.98



Valeurs significatives du coefficient de corrélation

v / α	0.10	0.05	0.02	v / α	0.10	0.05	0.02
1	0.9877	0.9969	0.9995	**16**	0.4000	0.4683	0.5425
2	0.9000	0.9500	0.980	**17**	0.3887	0.4555	0.5285
3	0.8054	0.8783	0.9343	**18**	0.3783	0.4438	0.5155
4	0.7293	0.8114	0.8822	**19**	0.3687	0.4329	0.5034
5	0.6694	0.7545	0.8329	**20**	0.3598	0.4227	0.4921
6	0.6215	0.7067	0.7887	**25**	0.3233	0.3809	0.4451
7	0.5822	0.6664	0.7498	**30**	0.2960	0.3494	0.4093
8	0.5494	0.6319	0.7155	**35**	0.2746	0.3246	0.3810
9	0.5214	0.6021	0.6851	**40**	0.2573	0.3044	0.3578
10	0.4973	0.5750	0.6581	**45**	0.2428	0.2875	0.3384
11	0.4762	0.5529	0.6339	**50**	0.2306	0.2732	0.3218
12	0.4575	0.5324	0.6120	**60**	0.2108	0.2500	0.2948
13	0.4409	0.5139	0.5923	**70**	0.1954	0.2319	0.2737
14	0.4259	0.4973	0.5742	**80**	0.1829	0.2172	0.2565
15	0.4124	0.4821	0.5577	**90**	0.1726	0.2050	0.2422
				100	0.1638	0.1946	0.2301

Corrélation ≠ causalité

Quelques exemples d'événements corrélés n'impliquant pas forcément une causalité :

- Le fait de dormir avec ses chaussures est fortement corrélé avec le fait de se réveiller avec la « gueule de bois ». Doit-on en conclure que dormir avec ses chaussures donne la « gueule de bois » ? Ou un troisième facteur est-il impliqué ?
- La fréquence des attaques de requins est fortement corrélée à la vente de glaces sur les plages ! Manger des glaces rend-il plus appétissant pour les requins ?
- Les personnes qui meurent ont très fréquemment vu un médecin dans les jours qui ont précédé. Est-il si dangereux de rencontrer un médecin ?
- Dans la plupart des villes, on constate une forte corrélation positive entre le nombre de nids de cigognes et la natalité. Nous cacherait-on des choses ?
- ...

Revue de presse

Pour être heureux, il faut faire plus souvent l'amour que la moyenne

Une chose est sûre : faire l'amour rend heureux ! Mais, d'après une étude publiée dans le magazine Social Indicators Research, il faudrait le faire plus souvent que les autres pour être encore plus content.

En effet, dans son étude sur "Le sexe et la poursuite du bonheur : Comment la vie sexuelle des autres est reliée à notre sentiment de bien être", le professeur de sociologie à l'université du Colorado explique que notre bonheur dépend de la fréquence de nos rapports sexuels. Elle devrait effectivement être plus élevée que la moyenne nationale. "Avoir plus de relations sexuelles augmente le bonheur, mais ce qui rend les gens encore plus heureux c'est de penser que notre fréquence de rapports est supérieure à celle des autres," a expliqué Tim Wadswort.

Pour arriver à cette conclusion, le professeur a analysé les données concernant la vie et le niveau de bonheur de 15 386 personnes entre 1993 et 2006. Ce dernier a également étudié les facteurs sociaux comme le situation matrimoniale, les revenus ou l'éducation.

Des résultats surprenants

"Il y a une inflexion globale du sentiment de bonheur qui augmente avec la fréquence de l'activité sexuelle" a déclaré Tim Wadswort.

Mais, on peut très bien se demander : « comment connaît-on la fréquence sexuelle de notre prochain? ». Le professeur Wadsworth avance que les gens se basent sur les médias, les études sur le sexe et même les émissions de télévision. "Il y a beaucoup de preuves que les informations concernant le comportement sexuel sont apprises via les groupes de références et les réseaux d'amitiés", explique-t-il.

La moyenne nationale

Et, si on se base sur la la moyenne nationale française, le nombre de rapports sexuels moyen par mois est de 8,9. Pour être parfaitement heureux, il suffirait donc de faire plus que ce chiffre pour se hisser au-dessus de la moyenne.

http://www.ladepeche.fr/article/2013/04/18/1608813-etre-heureux-faut-faire-plus-souvent-amour-moyenne.html
Publié le 18/04/2013 à 10:12, Mis à jour le 18/04/2013 à 14:34

Revue de presse

Faut-il manger du chocolat pour avoir des prix Nobel ?

©REUTERS/Denis Balibouse ©REUTERS/Denis Balibouse

La revue médicale New England Journal of Medicine vient de publier une étude qui fait le lien entre une forte consommation de chocolat et l'attribution des Nobel.
New England Journal of Medicine, hebdomadaire américain, publié depuis 1812, est considéré comme la revue médicale la plus prestigieuse.
Selon l'équipe de chercheurs, les flavonoïdes, de puissants antioxydants qu'on trouve en grande quantité dans les fèves de cacao, le thé vert et le vin rouge, ont montré qu'ils réduisent le risque de démence et améliorent les fonctions mentales chez les personnes âgées.

Le docteur Franz Messerli, de l'université Columbia à New York et auteur de l'étude, explique "qu'il y a une **corrélation significative surprenante** entre la consommation de chocolat per capita et le nombre de lauréats du Nobel pour dix millions d'habitants dans un total de 23 pays".
La Suisse arrive en tête à la fois en nombre de Nobel (rapporté à sa population) et en consommation de chocolat... Suivent ensuite, la France, l'Allemagne et les États-Unis.
Inversement, la Chine, le Japon et le Brésil se montrent moins gourmands en chocolat et aussi en Nobel.
Seule exception de l'étude : la Suède. Les habitants ne consomment "que" 6,4 kilos de chocolat par an et par personne pour un total de 32 Nobel. Qu'à cela ne tienne, pour les chercheurs il s'agirait d'un simple favoritisme du comité Nobel.
Si une corrélation est montrée pour les pays, l'étude ne dit rien en revanche sur le niveau de consommation individuel de chocolat des lauréats du Nobel.

Revue de presse

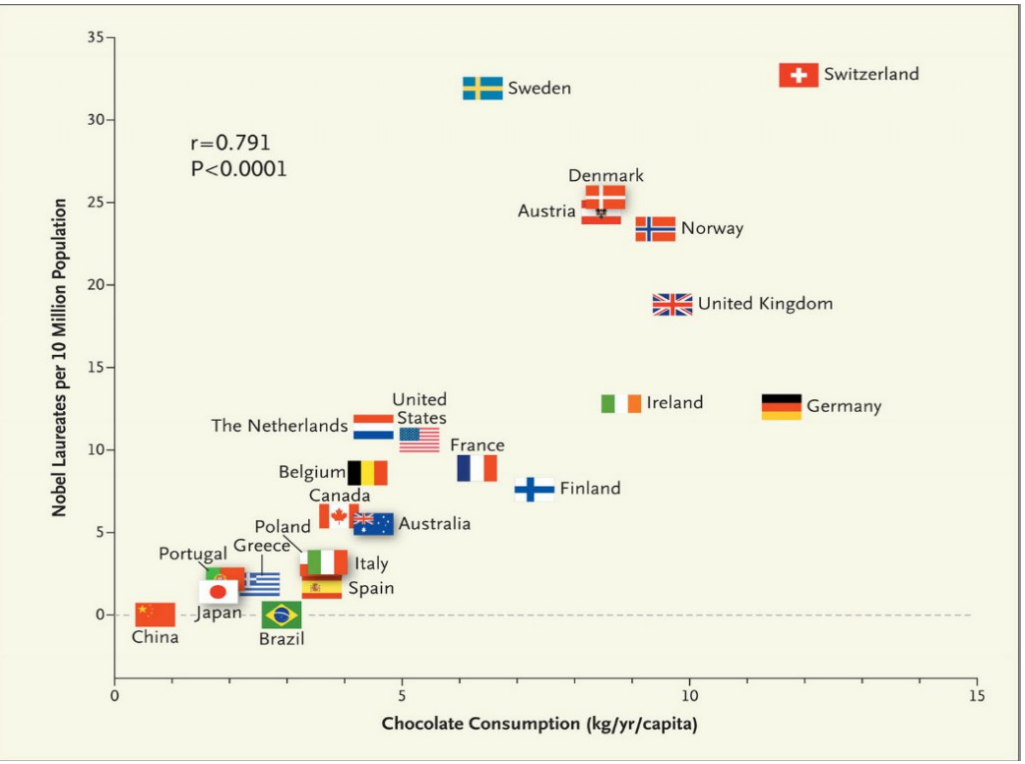
Le chocolat engendre des tueurs en série

http://plus.lefigaro.fr/lien/le-chocolat-engendre-des-tueurs-en-serie-20121123-1589103

23/11/2012 | Mise à jour: 11:12 Réactions (24)
Sélectionné par la rédaction
Par Hayat Gazzane

Des chercheurs britanniques se sont amusés à démonter l'étude de Franz Messerli qui établit une corrélation forte entre consommation de chocolat et prix Nobel. En utilisant la même méthodologie, ils arrivent à prouver que les pays où l'on mange beaucoup de chocolat sont aussi ceux qui engendrent le plus de sérial killer et d'accidents de la route (étude en anglais).

Merci pour le chocolat ?



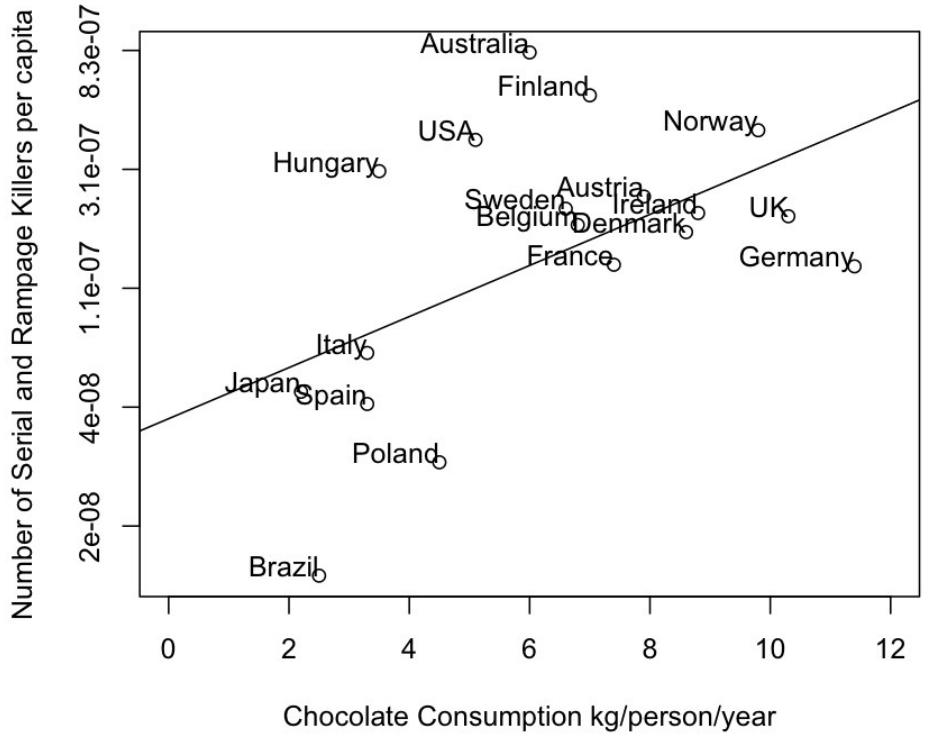


Figure 1: Correlation between countries' annual per capita chocolate consumption and the serial and rampage killers per capita since 1900.

Revue de presse

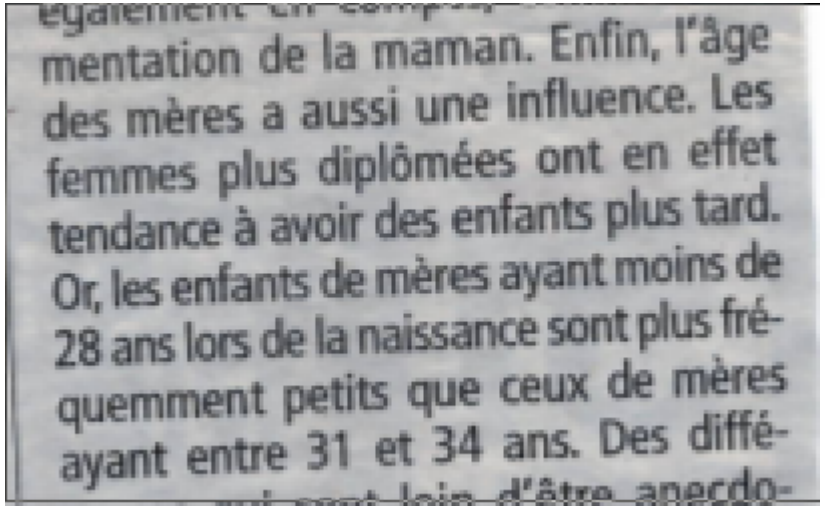
14 N° 1278 JEUDI 11 JUIN 2015 France

ÉTUDE SUR LA MATERNITÉ

PAS DE DIPLÔME, DE PETITS BÉBÉS

Selon l'Ined, les mères peu instruites auraient des bébés plus petits que celles diplômées.

Le niveau d'études de la mère aurait une influence sur le poids du bébé. C'est ce que révèle une étude de l'Ined publiée hier : moins une femme est diplômée, plus elle risque de mettre au monde un enfant de faible constitution (moins de 2,5 kg à la naissance). Pour les mamans n'ayant aucune formation, le risque est 50 % plus élevé qu'elles aient un enfant petit, comparé à celles ayant le baccalauréat. Et celles qui n'ont que le bac ont 25 % plus de risques que celles qui ont fait plus de deux ans d'études. Un lien qui pourrait s'expliquer, selon les chercheurs, par le fait que les femmes moins instruites ont davantage tendance à fumer pendant la grossesse, ce qui a des conséquences sur le développement du fœtus. D'autres facteurs psychosociaux entreraient également en compte, comme l'alimentation de la maman. Enfin, l'âge des mères a aussi une influence. Les femmes plus diplômées ont en effet tendance à avoir des enfants plus tard. Or, les enfants de mères ayant moins de 28 ans lors de la naissance sont plus fréquemment petits que ceux de mères ayant entre 31 et 34 ans. Des différences qui sont loin d'être anecdotiques : le poids du bébé est en effet un indicateur de santé à la naissance. •



mentation de la maman. Enfin, l'âge
des mères a aussi une influence. Les
femmes plus diplômées ont en effet
tendance à avoir des enfants plus tard.
Or, les enfants de mères ayant moins de
28 ans lors de la naissance sont plus fré-
quemment petits que ceux de mères
ayant entre 31 et 34 ans. Des diffé-

8 metronews AU QUOTIDIEN #sciences mardi 2 juin 2015 www.metronews.fr

2 sur metronews.fr

SOCIAL Le gouvernement annonce une nouvelle série de mesures pour simplifier les démarches administratives. *Metronews* vous explique les 5 points qui vont changer.

DIAPORAMA

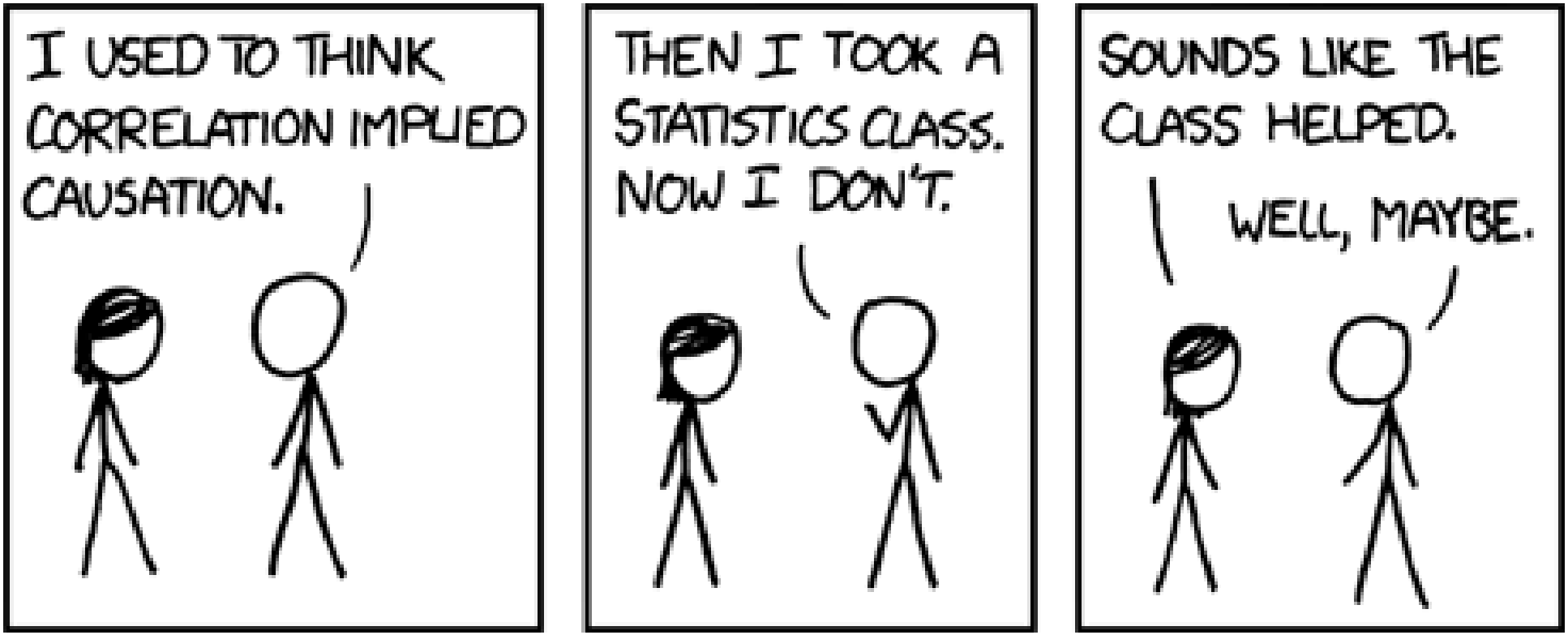
Le télescope Hubble a photographié Nasty 1, une étoile massive au vieillissement accéléré. *Metronews* vous fait découvrir ses premiers clichés.

Bien-être

Le bruit aussi fait grossir

Sans pour autant y voir un lien de cause à effet, les chercheurs indiquent que l'accumulation de graisse peut être due au cortisol : une hormone stéroïde libérée dans l'organisme en réponse à un stress physique ou psychologique. Cette hormone joue un rôle déterminant pour réguler la tension artérielle et les fonctions cardiovasculaire et immunitaire. De plus, en nuisant au sommeil, l'excès de bruit peut aussi affecter notre métabolisme et ainsi augmenter la prise de poids. •

Humour de statisticien...



Source : xkcd.com

Corrélation & causalité

http://plus.maths.org/content/coincidence-correlation-and-chance

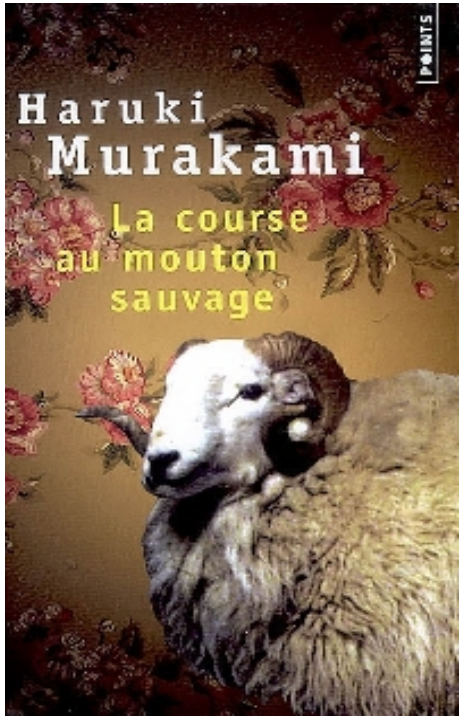
And talking of Jenny, there she is now, buying an ice cream from the local shop. With her family about to go out to Australia for a holiday, I ought to go and warn her that **the more ice creams there are sold, the more shark attacks there are**. Again, I've done my research quite thoroughly, and the numbers do not lie. Perhaps I should recommend an apple instead!

Finally, let's pop into my local primary school to chat to the head teacher. I want to tell her about research I've uncovered which shows a clear and **proven link between literacy levels and hand size in children.** Bigger hands make better readers, it seems. With my son starting there in the autumn, maybe now is the time to set up some sort of hand-stretching programme - perhaps on Wednesday afternoons, now that PE's been scrapped?

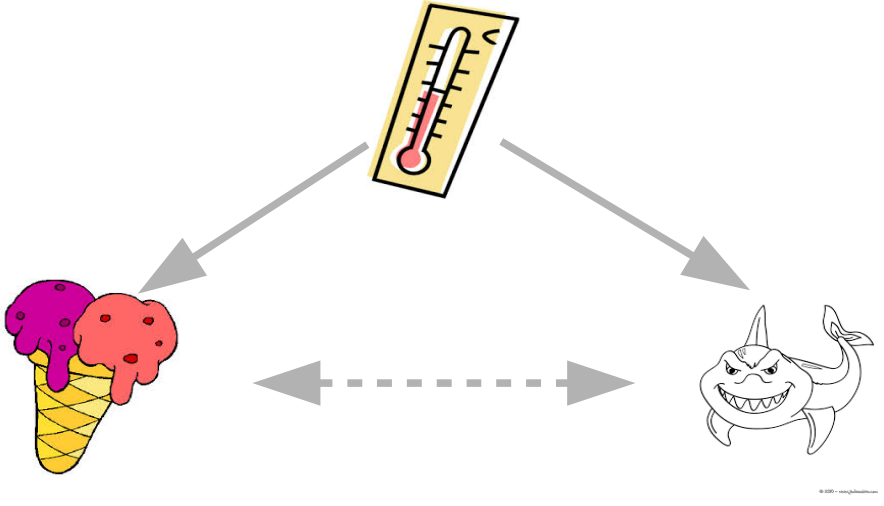
These examples may seem bizarre and improbable, but they are not the result of bad statistics. All the information is absolutely correct. **Their strangeness comes from our own reasoning. We see two things changing together and our instinct is to assume that they are tied by cause and effect**. Unfortunately, our instinct is often wrong. In all these examples **a third "confounding" variable is actually the cause of two correlated variables**.

It is absolutely true that **people who play loud music are more likely to suffer from acne, but only because teenagers make up a big part of both groups**. Acne and loud music are certainly correlated. **But correlation is not causation**. The same thing is true with the sharks and ice cream. The number of shark attacks and ice creams sold both go up during the summer, with the good weather encouraging people both to go in swimming and to eat ice cream. And as for large hands? Older children are bigger, and can read better!

Corrélation ≠ causalité



Votre hypothèse à vous, dis-je, serait qu'entre ces deux phénomènes il n'y aurait pas de relation de cause à effet mais une simple situation de parallélisme derrière laquelle il y aurait un autre et mystérieux facteur.

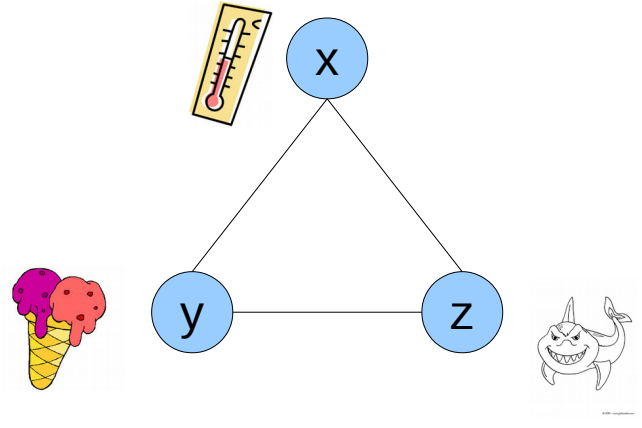


Corrélation partielle

Formellement, la corrélation partielle entre X et Y conditionnellement à Z est la corrélation entre les résidus RX et RY des régressions linéaires de X sur Z et de Y sur Z.

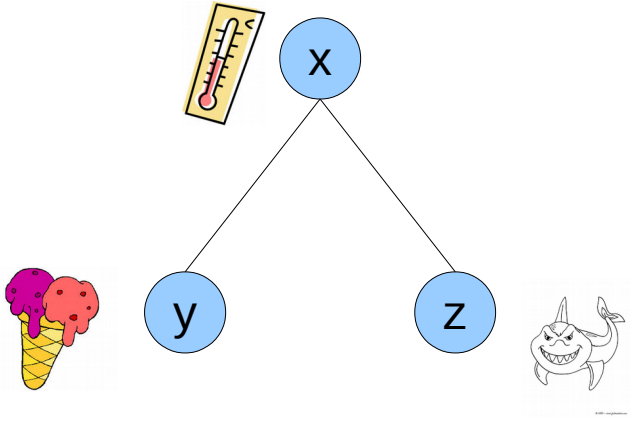
Matrice de corrélation

	x	y	z
x	1.00	**0.95**	**0.93**
y	.	1.00	**0.88**
Z	.	.	1.00



Matrice des corrélations partielles

	x	y	z
x	1.00	**0.74**	**0.66**
Y	.	1.00	**0.04**
z	.	.	1.00



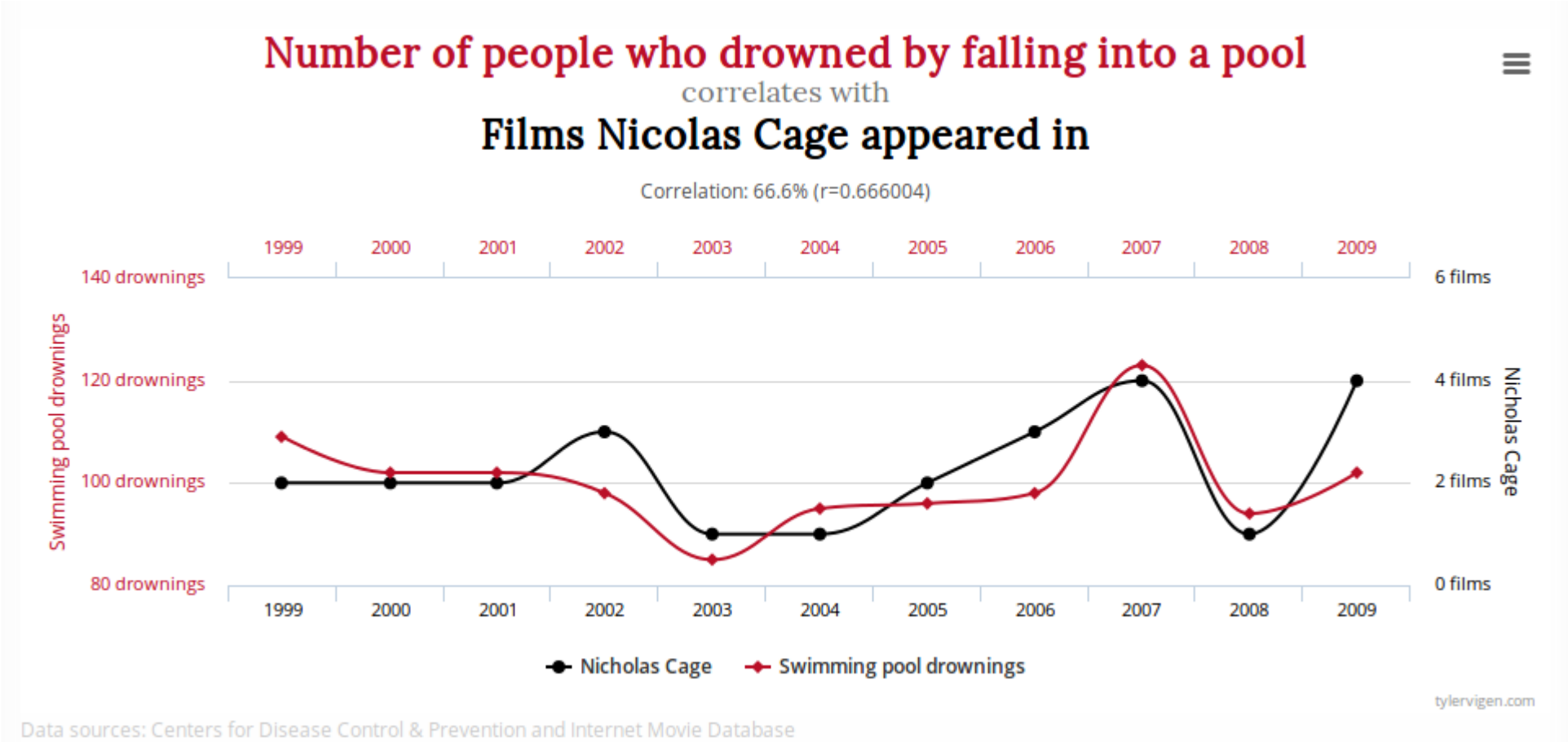
Corrélation partielle

Pr Jacques Baillargeon
http://www.uqtr.uquebec.ca/~baillarg/srp-6001/cours3/partielle.htm

Il est surprenant de constater qu'une technique statistique aussi puissante et aussi facile à obtenir que la corrélation partielle ne soit pas plus fréquemment utilisée en psychologie. Cette technique permet d'évaluer la corrélation entre deux variables après avoir contrôlé l'effet perturbateur d'une ou de plusieurs autres variables.

Spurious correlation

`http://www.tylervigen.com/spurious-correlations`



Divorce rate in Maine / Consumption of margarine (ρ= 0.99)
Consomption of Mozarelle cheese / Civil engineering doctorates awarded (ρ= 0.96)
...

Représentations graphiques

Représenter une ou plusieurs séries de chiffres par un graphique

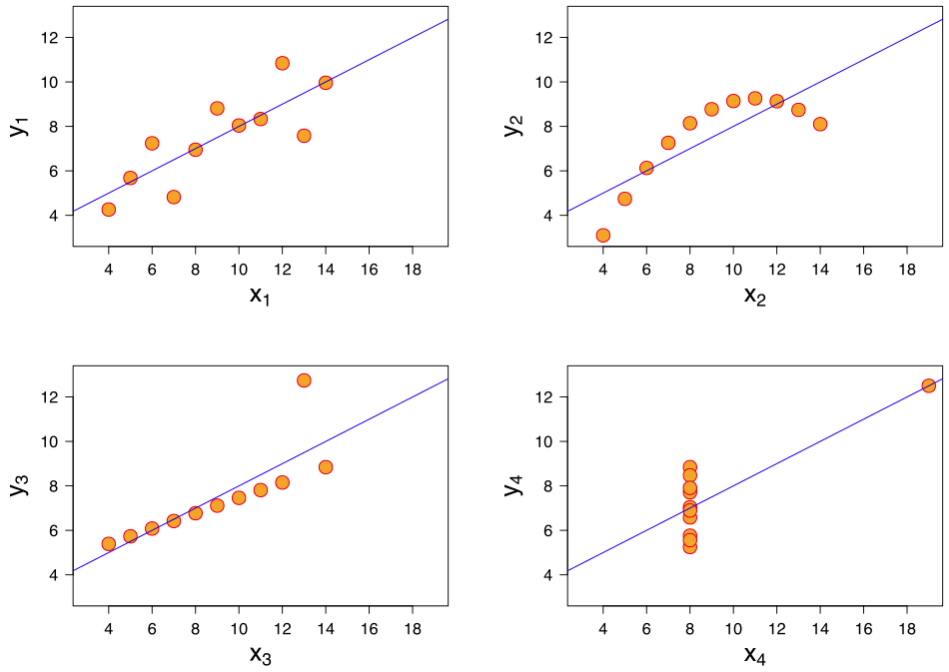
Pourquoi visualiser ?

...make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding.

F. J. Anscombe, 1973

Anscombe's quartet

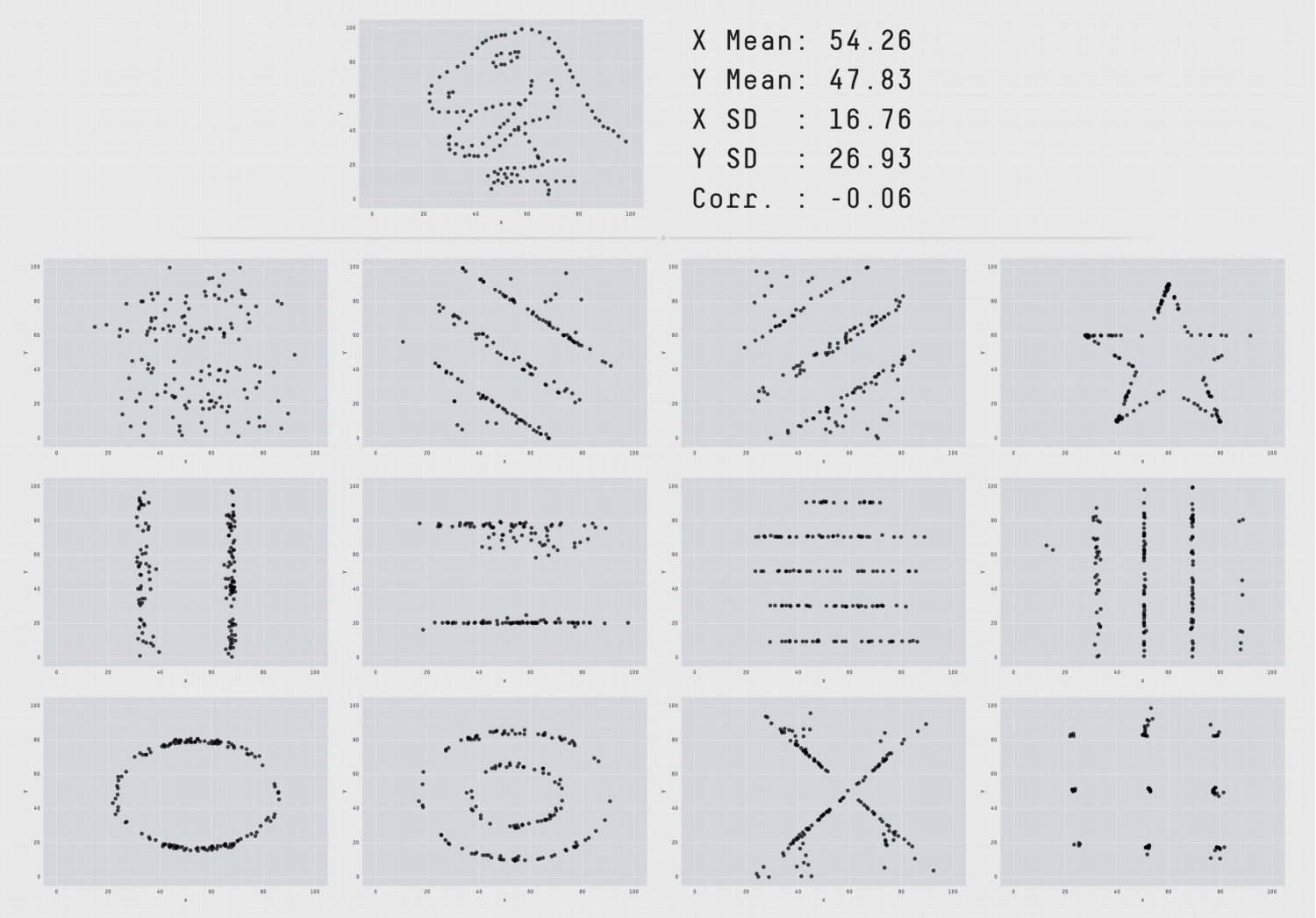
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



Dans les 4 cas :

	X	Y
Moyenne	9	7.5
Variance	11	4.125
Corrélation	0.816	

Same stats, different graphs...



Matejka, J., & Fitzmaurice, G. (2017, May). **Same stats, different graphs**: Generating datasets with varied appearance and identical statistics through simulated annealing. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 1290-1294). ACM.

www.autodeskresearch.com/publications/samestats

Un peu de théorie

Un graphique devrait :

- Montrer les données
- Inciter celui qui regarde à penser
- Éviter de distordre ce que les données ont à dire
- Présenter beaucoup de données sur une petite surface
- Révéler les données à des niveaux différents : d'un aperçu global à des structures plus fines
- Servir un objectif clair et raisonnable
- Être étroitement intégré à une description statistique du jeu de données

Edward Tufte, *The Visual Display of Quantitative Information*, Cheshire, CT, Graphics Press, 2001, 2e éd. (1re éd. 1983)

Représentations graphiques

Données de type « effectif »

A	30
B	15
C	30
D	20
E	25

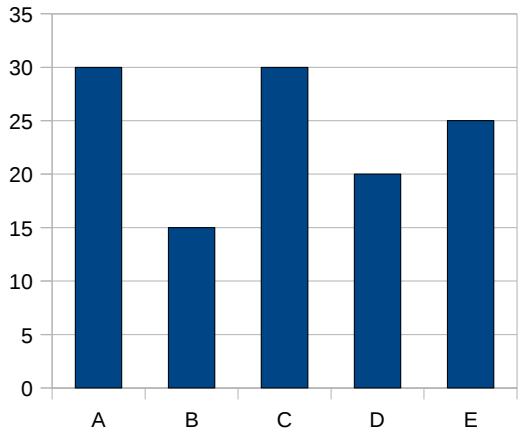
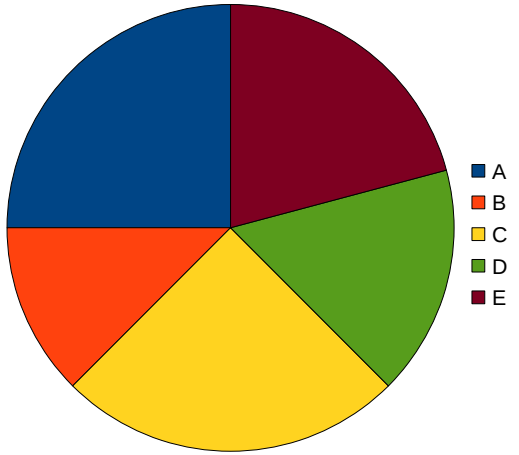
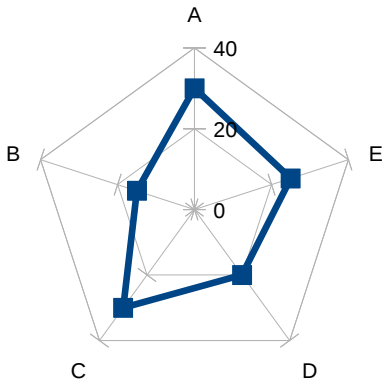


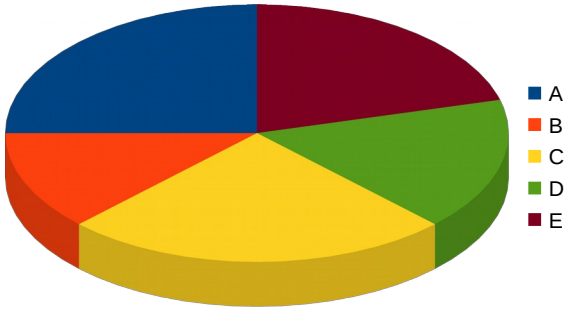
Diagramme en bâtons

Diagramme circulaire



Spiderman plot :-)

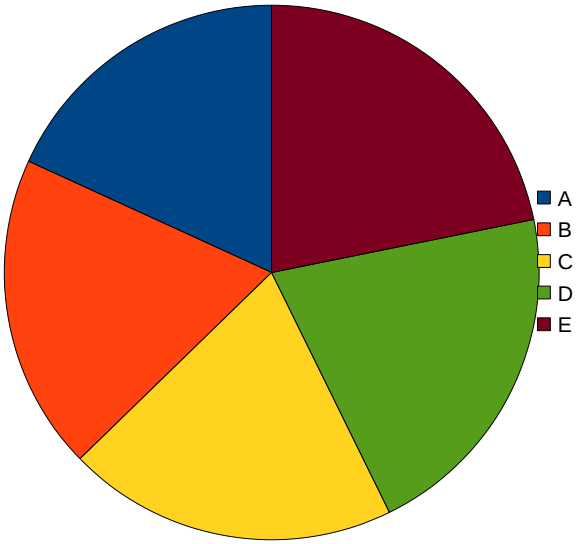




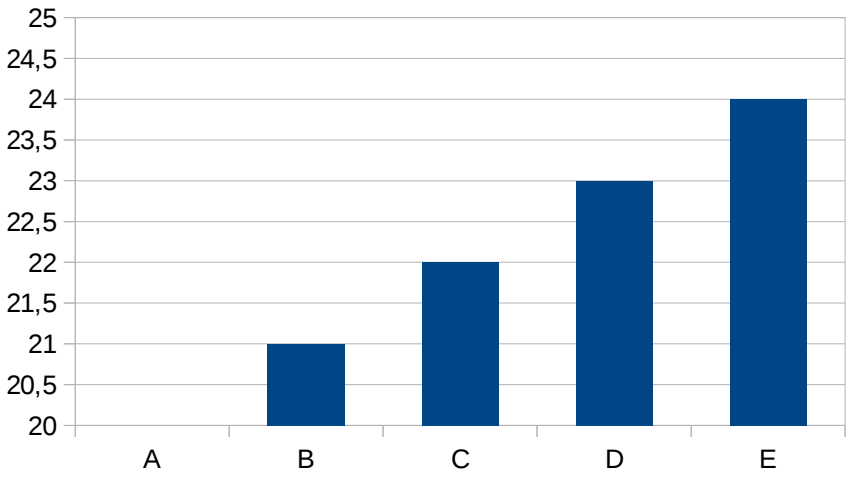
Attention aux représentations 3D. Est-ce si évident que A et C ont la même valeur ?

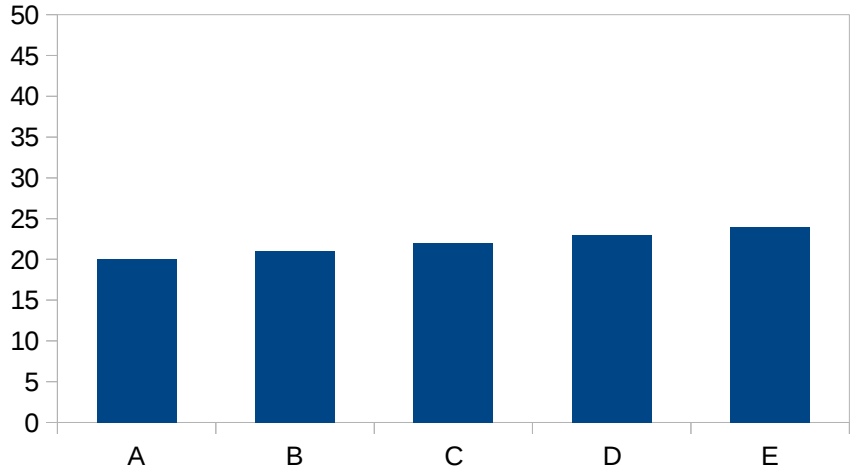
Représentations graphiques

Données de type « effectif »

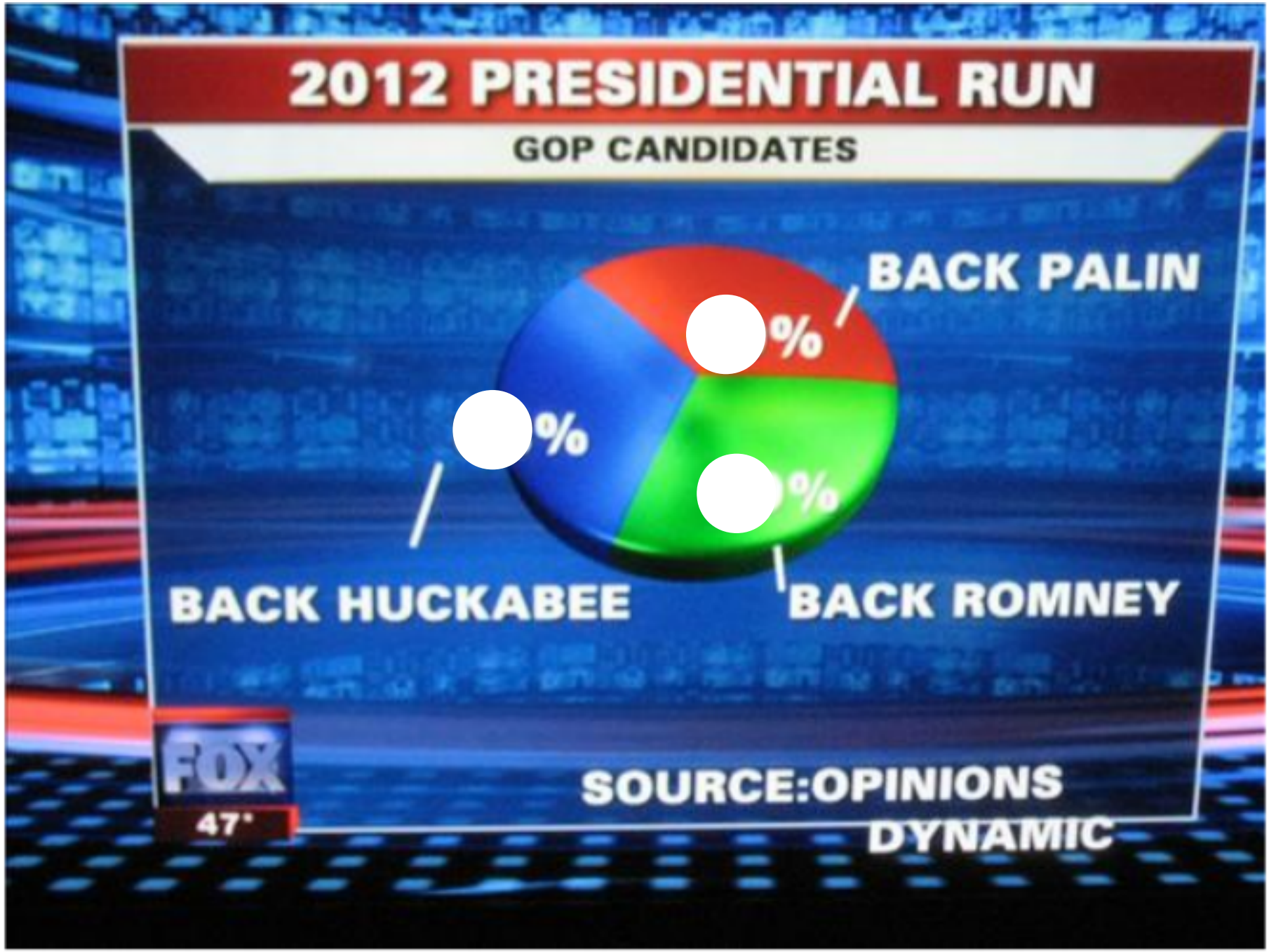


A	20
B	21
C	22
D	23
E	24

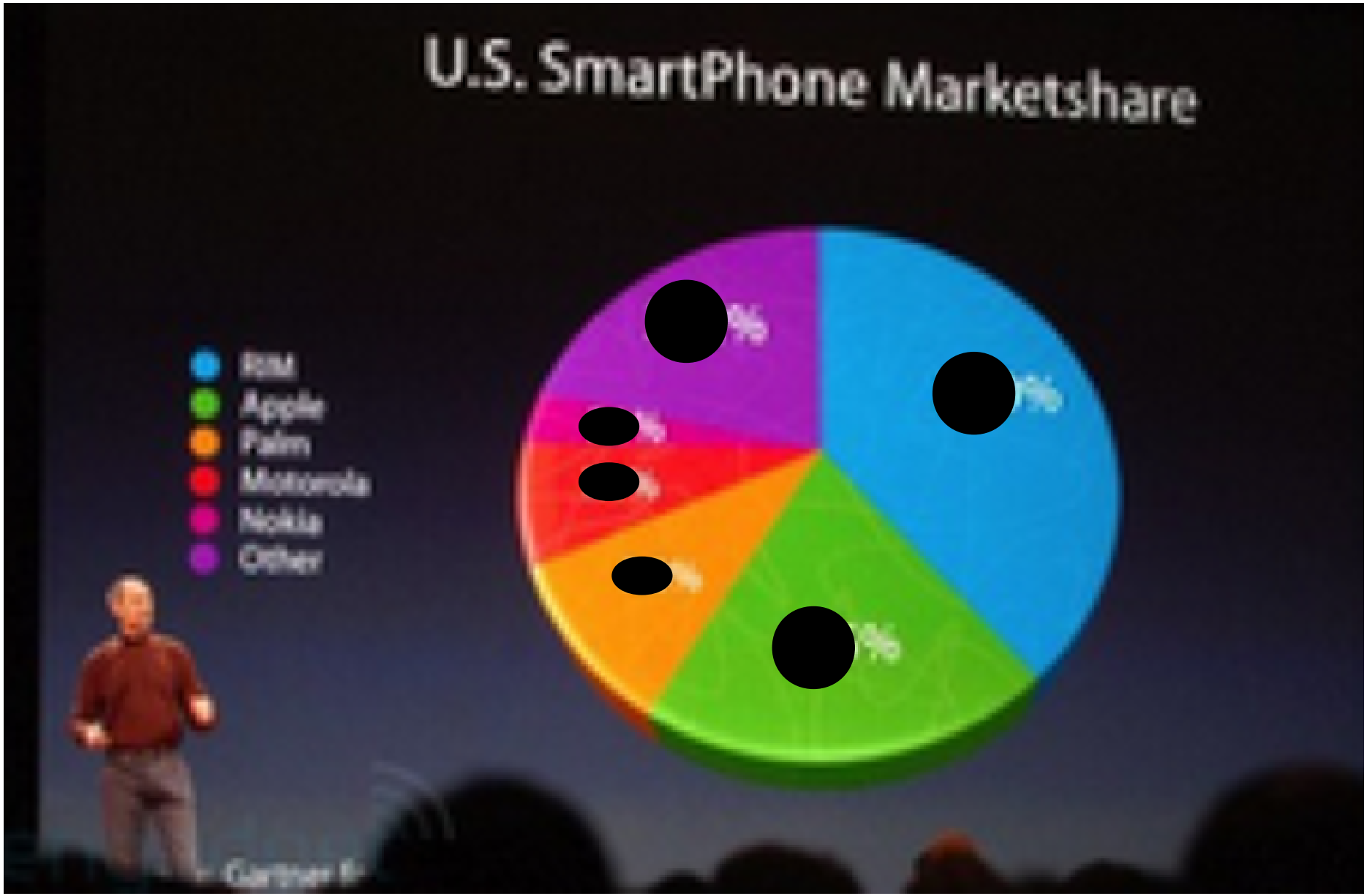




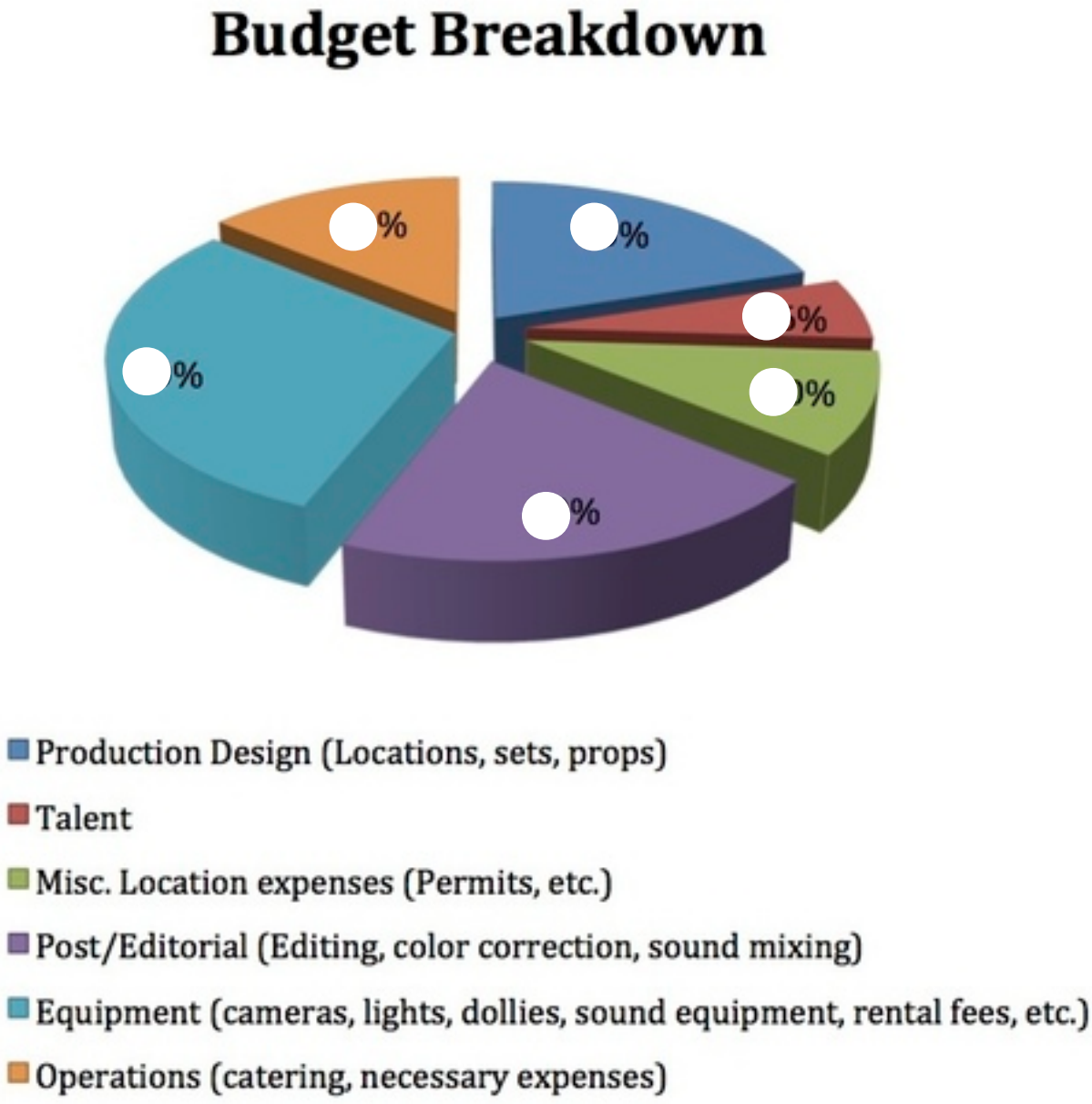
Hum, hum...


2012 PRESIDENTIAL RUN
GOP CANDIDATES
BACK PALIN
%
%
%
BACK HUCKABEE
BACK ROMNEY
FOX
47°
SOURCE:OPINIONS
DYNAMIC

Hum, hum...

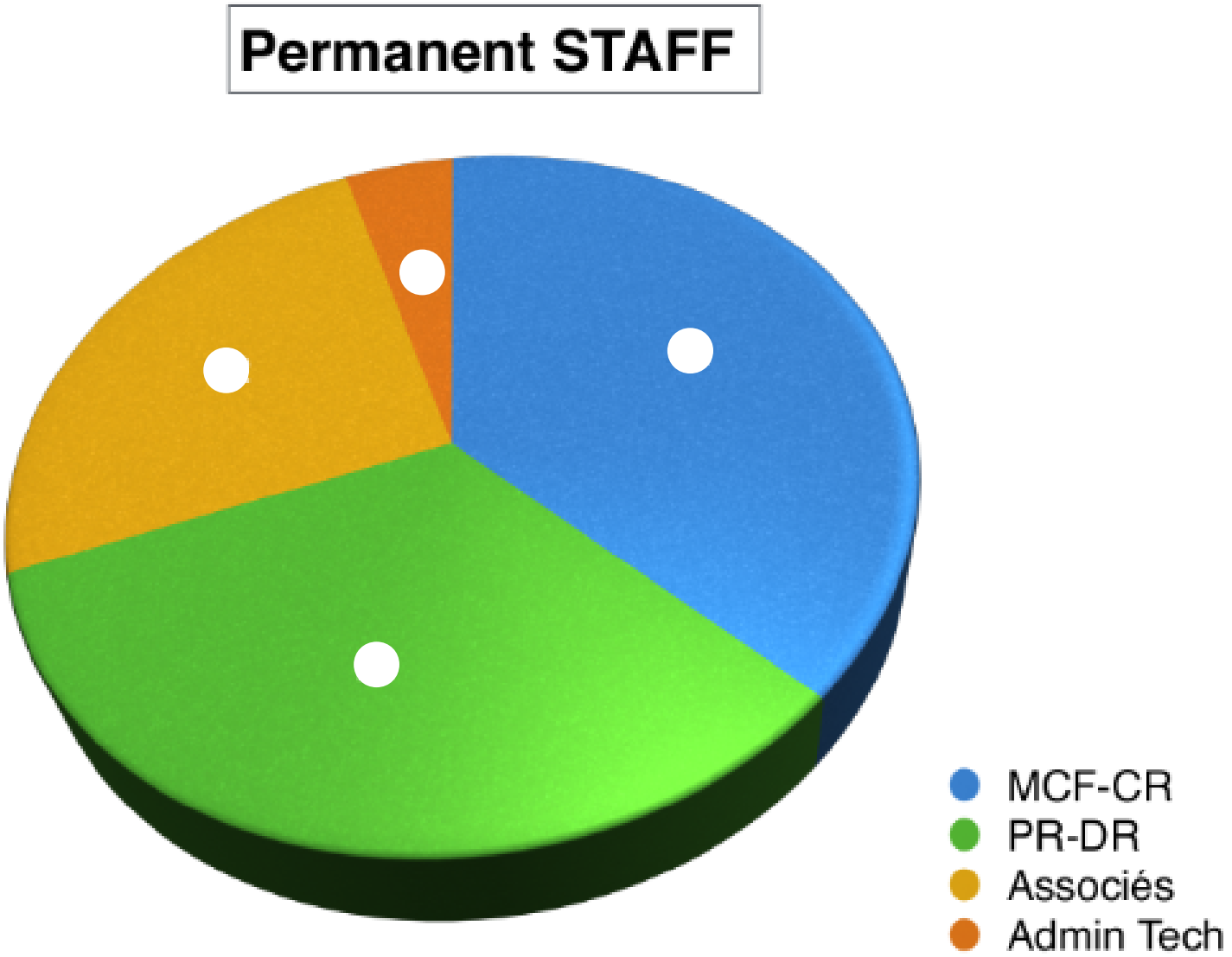

U.S. SmartPhone Marketshare
RIM
Apple
Palm
Motorola
Nokia
Other

Hum, hum...


Budget Breakdown
%
%
%
%
%
Production Design (Locations, sets, props)
Talent
Misc. Location expenses (Permits, etc.)
Post/Editorial (Editing, color correction, sound mixing)
Equipment (cameras, lights, dollies, sound equipment, rental fees, etc.)
Operations (catering, necessary expenses)

Hum, hum...

ESP in few facts : 88 Permanent people



Revue de presse

VENDREDI 5 JUIN 2009 EUROPE 11

SONDAGE Ipsos dévoile les dernières intentions de vote pour « 20 Minutes »-SFR-« Le Point »

LE SCRUTIN DE DIMANCHE, UN TOUR DE CHAUFFE ?

BASTIEN BONNEFOUS

Qui sortira en tête dimanche soir ? Le PS finira-t-il 2e ? A qui la 3e place : au centre ou aux Verts ? Derniers points d'étape, d'après les intentions de vote Ipsos pour *20 Minutes*-SFR-*Le Point*.

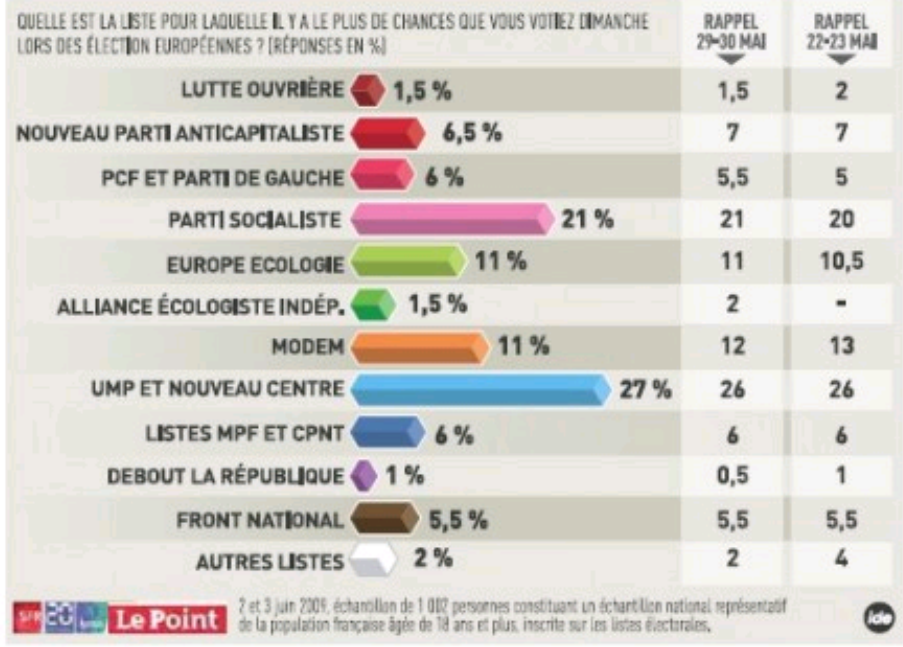
◗ **UMP-Majorité parlementaire** Plus vraiment de doute, le parti présidentiel devrait finir en tête du scrutin, avec quelque 27 % prévus. L'écart avec le PS se creuse de six points, mais, surtout, 85 % de l'électorat UMP dit avoir fait définitivement son choix. Un socle solide, qui permettrait ainsi à l'UMP de passer la barre des 25 % qu'il s'est lui-même fixée. Serait-ce pour autant une victoire ? « Elle serait à l'image de la popularité de Nicolas Sarkozy : solide chez ses partisans, mais qui a du mal à s'élargir. C'est suffisant pour un scrutin à un tour comme les européennes, mais cela pose des problèmes de réserves de voix dans des votes à deux tours », estime Jean-François Doridot, directeur général d'Ipsos. Message reçu pour les régionales 2010 et la présidentielle 2012.

◗ **PS** Avec 21 % d'intentions de vote, les socialistes dépassent à peine le niveau clé des 20 %. Pas sûr même, selon les sondeurs, que le PS y parvienne. Un mauvais score aurait des conséquences lourdes dès dimanche soir : une relance illico des divisions internes jamais éteintes dans le parti, un allié Verts de plus en plus gourmand, et le risque que le PS ne soit plus majoritaire à terme au sein de la gauche.

◗ **La 3e place** Impossible de départager le MoDem et Europe Ecologie (11 % chacun). « La dynamique de fin de campagne profite aux Verts, qui bénéficient d'une image de vote protestataire plus utile que le NPA », estime Jean-François Doridot. Par sa campagne très antisarkozyste, le MoDem pourrait avoir crispé son électorat de droite et celui, europhile, de l'ex-UDF, au profit de l'UMP et des Verts.

◗ **La gauche radicale** Même situation du côté du NPA et du Front de gauche, au coude à coude autour de 6 %. Un score en dessous des espoirs de début de campagne portés par le discours sur la crise. « Le NPA, qui touche un électorat jeune, donc volatil, n'a pas réussi à convaincre d'un vote utile ; le Front de gauche a marqué des points, profitant de l'électorat communiste, structuré, âgé et mobilisé », estime Jean-François Doridot. ■

L'UMP TOUJOURS EN TÊTE



■ EXTRÊME DROITE

Les listes FN et Libertas de Philippe de Villiers gravitent autour de 6 %, alors que ce camp a toujours profité par le passé du scrutin européen. « Leur thème classique du protectionnisme a été repris par la plupart des autres partis, de gauche comme de droite », explique Ipsos.

Revue de presse

Développement

TPE/PME

Ardan réveille les projets des petites entreprises

ans les petites entreprises, le dirigeant umule souvent plusieurs casquettes : esponsable financier, de la production, es ressources humaines, responsable ommercial, marketing et juridique, etc. our faire grandir l'entreprise - qu'il 'agisse de lancer un nouveau produit, le démarcher de nouveaux clients… il lui faut souvent se résoudre à onfier ce projet à une personne édiée. Une étape difficile à franchir, ui fait souvent hésiter les patrons de etites entreprises : ce nouveau salarié a-t-il me permettre de faire augmenter e chiffre d'affaires de l'entreprise ? Comment s'intégrera-t-il ?

Répartition des dispositifs par fonction

Pour aider les chefs d'entreprises à sauter le pas sans mettre en péril leur entreprise, il existe le dispositif ARDAN Développeur, mis en place par le CNAM (Centre National des Arts et Métiers) et soutenu financièrement par la Région Midi-Pyrénées. Le principe est simple. Pendant une période de 6 mois, le chef d'entreprise accueille un demandeur d'emploi ayant les compétences requises pour mener à bien son projet. Cela permet à la fois à celui-ci d'avoir le temps de faire ses preuves et au chef d'entreprise de s'assurer que son projet est viable.

Pendant ce stage, le demandeur d'emploi bénéficie également d'une formation qualifiante tout en conservant ses droits et ses indemnités chômage.

325 entreprises de la région ont bénéficié du dispositif depuis 2007.

La mise en place d'Ardan Développeur coûte au total 5 000 euros à l'entreprise, la Région prenant en charge le solde (5 500 euros).

Selon une étude réalisée par le CNAM Midi-Pyrénées en février dernier, 80 % des demandeurs d'emploi sont embauchés par l'entreprise à l'issue de ce « stage » de 6 mois dont 64 % en CDI. Ces embauches en entraînent également d'autres, en moyenne 1,3 emploi est créé en plus de celui du pilote du projet.

Autre particularité du dispositif, il permet de valoriser l'expérience des demandeurs d'emploi : un tiers des bénéficiaires du dispositif ont plus de 50 ans.

Du côté de l'entreprise, les bénéfices sont également visibles. 40 % des chefs d'entreprise déclarent avoir vu leur chiffre d'affaires augmenter et 43 % d'entre eux en ont profité pour investir dans de nouveaux projets.

Depuis janvier 2007, 325 entreprises (86 % d'entreprises de moins de 20 salariés) de Midi-Pyrénées ont bénéficié du dispositif Ardan.

Plus d'infos sur www.ipst-cnam.fr/Dipostif-Ardan-Developpeur

Les projets développés ont permis…

Typologie des projets

CONCOURS

34e édition des Inn'Ovations

Votre entreprise innove en Midi-Pyrénées ? Faites-le savoir et participez à la 34e édition des Inn'Ovations ! Vous avez jusqu'au 18 octobre prochain pour déposer votre candidature.

Les innovations
LE CONCOURS DE L'INNOVATION EN MIDI-PYRÉNÉES

Organisé par MPI, l'agence régionale de l'innovation et financé par la Région Midi-Pyrénées en partenariat avec Airbus, EDF et BNP Paribas, ce concours a pour objectif de récompenser et de soutenir les entreprises régionales innovantes. Histoire de ne pas passer à côté d'une pépite et de saluer tous les talents, le concours doté de 190 000 euros de prix comporte six catégories* dont trois nouvelles ainsi que deux prix spéciaux. Pour chacun de ces prix, le jury délibérera en tenant compte du caractère de l'innovation, de son originalité ou encore du champ d'application qu'elle couvre. Il prendra en considération l'existence d'une première démonstration, de la stratégie d'accès au marché mise en place, des retombées économiques envisagées, des passerelles de coopération induites entre les milieux dits « académiques » et économiques et enfin de la pertinence de la candidature par rapport à la catégorie dans laquelle l'entreprise s'est inscrite. La sélection aura lieu du 17 au 25 novembre 2014. Quant à la cérémonie de remise des trophées, celle-ci se déroulera le 29 janvier 2015, devant un large public, en clôture du salon Midinnov organisé chaque année par MPI pour la Région Midi-Pyrénées.

*Innovation, produits et services du futur/ Innovation, croissance et développement territorial/Innovation et société/Innovation et formation/ Innovation et international/ Innovation et jeunes entreprises/ Grand prix/ Prix coup de cœur.

Pour plus d'info, rendez-vous sur : www.mp-i.fr

Représentations graphiques

Données de type « effectif »

	X	Y
A	30	25
B	15	15
C	30	20
D	20	30
E	25	35

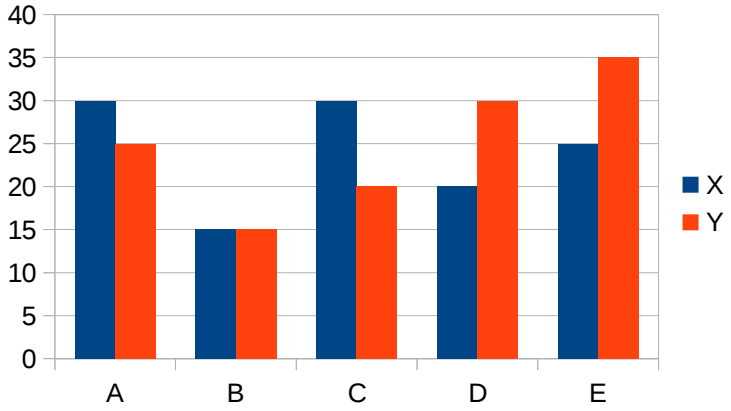
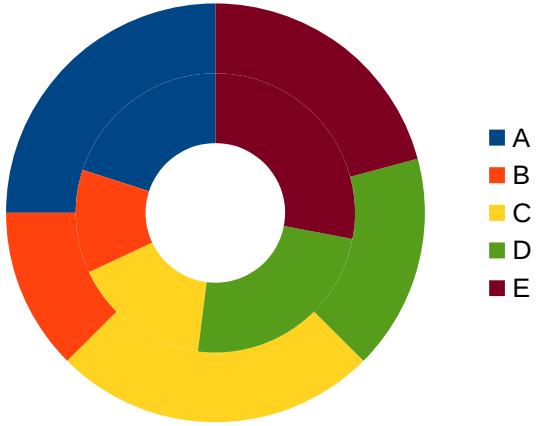
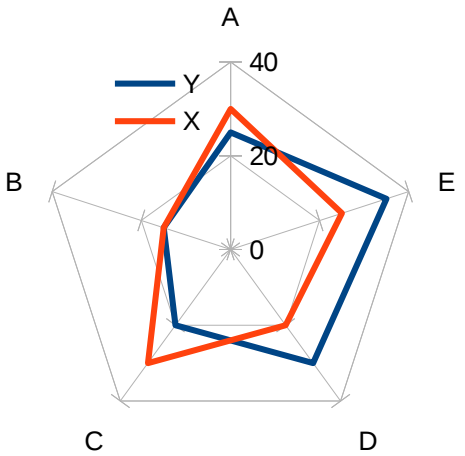


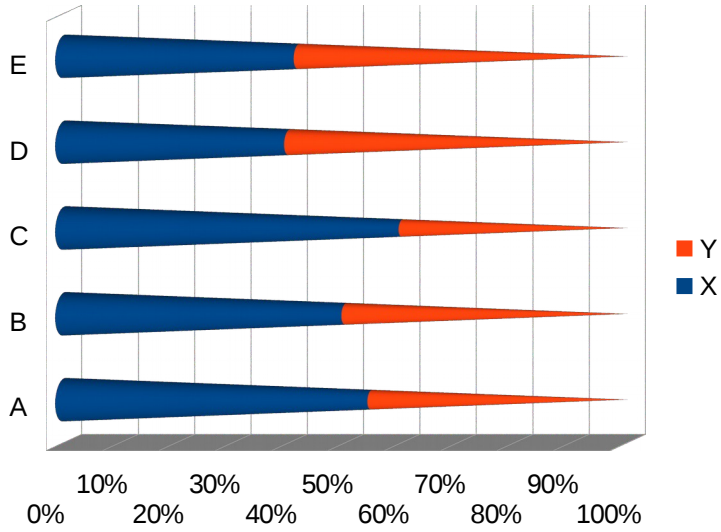
Diagramme en bâtons

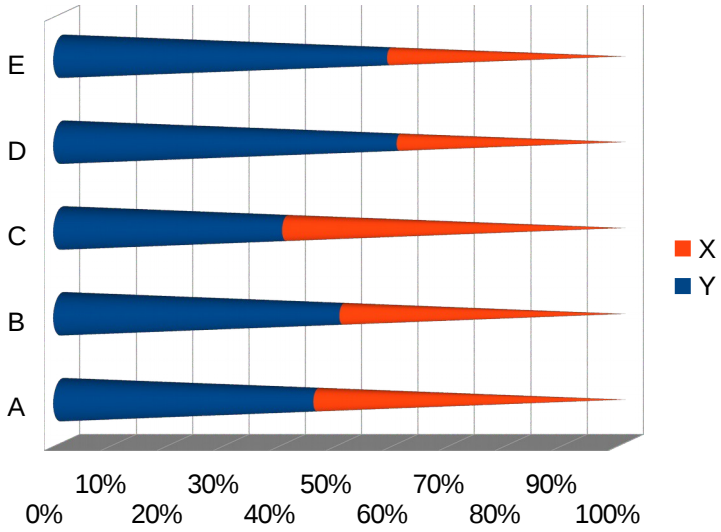
Diagramme « tore »





Cônes 3D (!?)





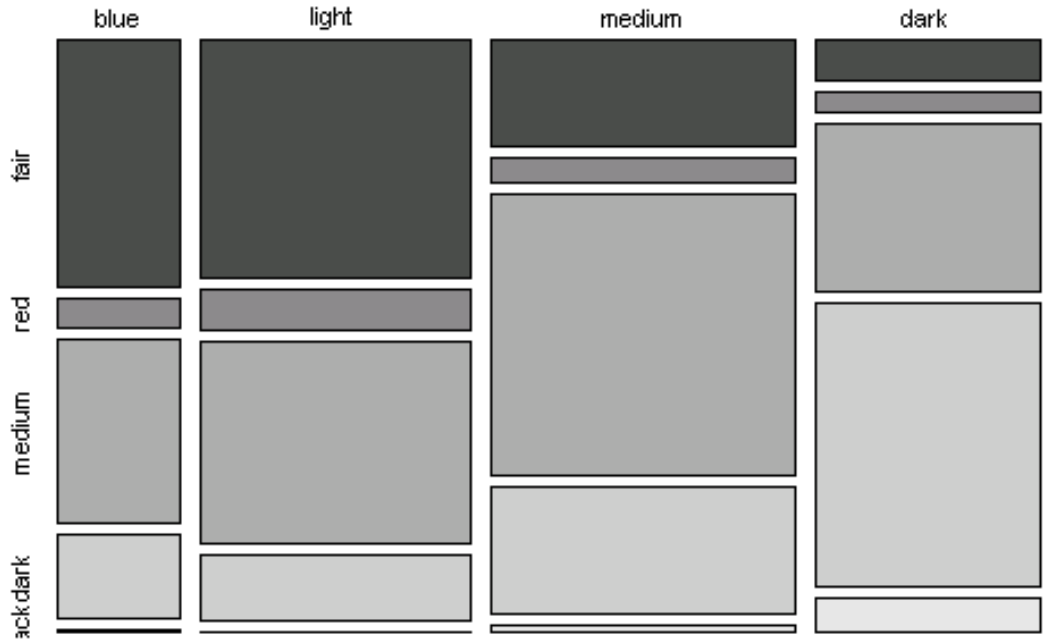
Représentations graphiques

Données de type « effectif » - Table de contingence

(données *caith*, R)

	fair	red	medium	dark	black
blue	326	38	241	110	3
light	688	116	584	188	4
medium	343	84	909	412	26
dark	98	48	403	681	85

Diagramme mosaïque



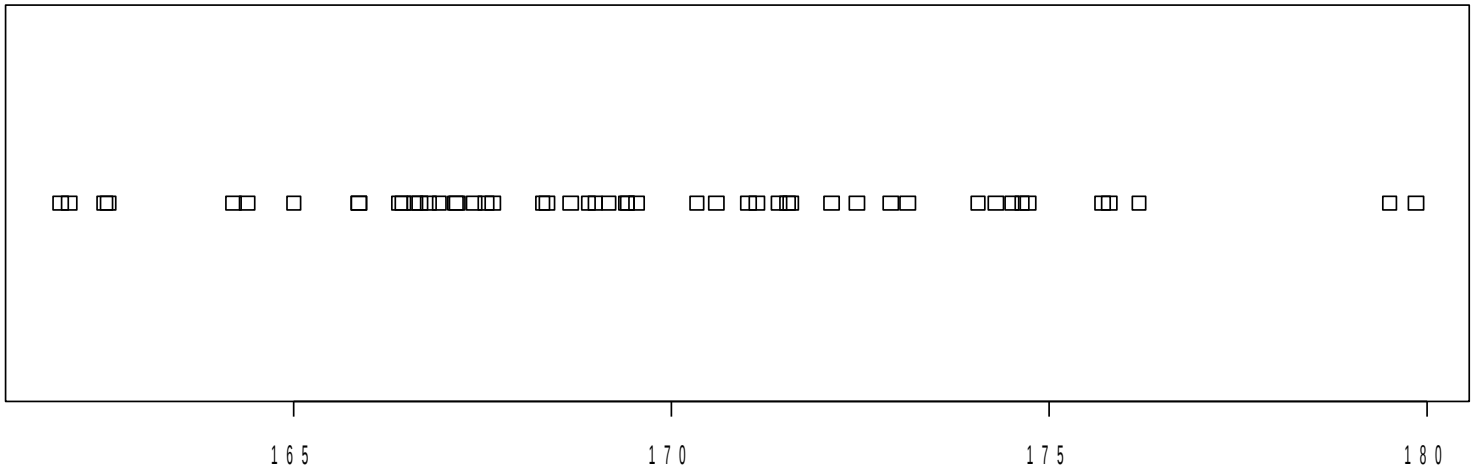
Représentations graphiques

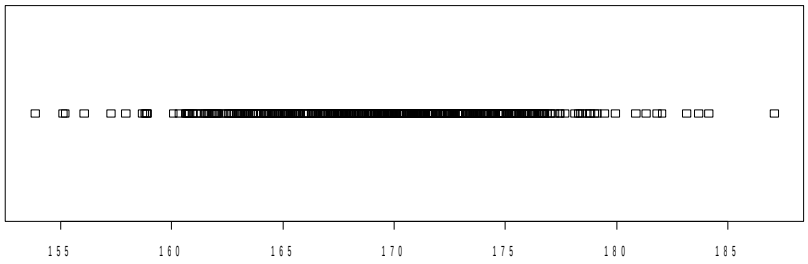
Données quantitatives

50 observations : taille en cm de 50 individus

171.4 175.8 168.9 166.8 174.3 169.2 165.9 171.6 167.1 179.8
174.5 169.0 170.6 173.1 169.4 167.4 162.5 172.9 174.7 164.2
175.7 167.2 169.5 174.1 162.5 172.4 166.4 162.0 167.5 168.7
166.9 171.0 176.2 172.1 166.4 168.3 170.3 164.4 167.6 168.3
165.9 174.6 171.5 169.4 166.6 179.5 166.7 171.1 161.9 165.0

Nuage de points 1D
(*stripchart*)





Avec 500 observations, la lisibilité devient délicate.

Boxplot

Q3 + 1.5(Q3-Q1) ou maximum

Q3 : 3ème quartile (75%)

Q2 : Médiane (50%)

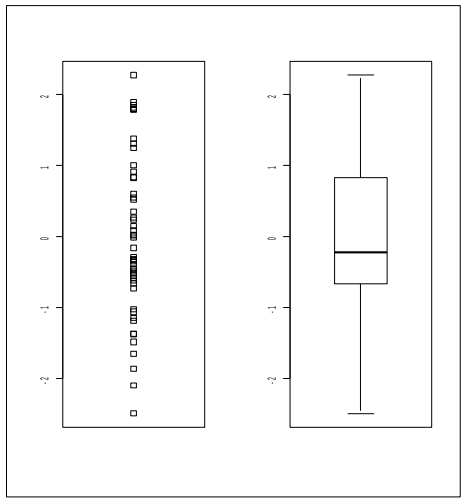
Q1 : 1er quartile (25%)

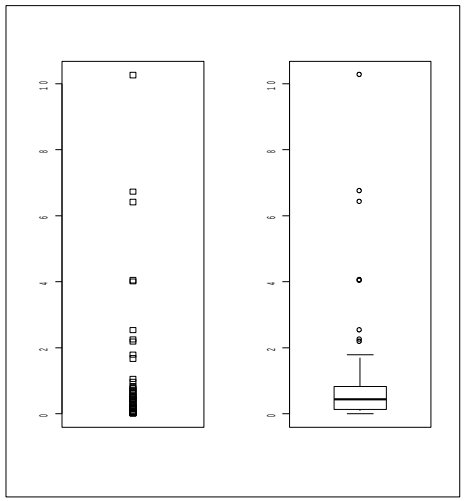
Espace inter-quartile

Q3-Q1

Q1 – 1.5(Q3-Q1) ou minimum

Points extrêmes * *





Boxplot : on dit aussi boîte à moustaches

Source : ML

Histogramme

≠

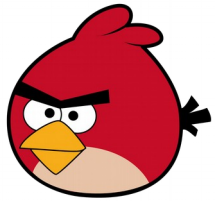
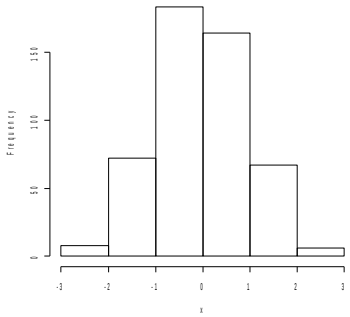
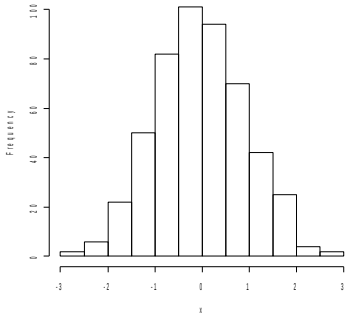
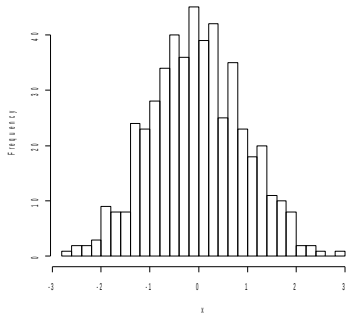
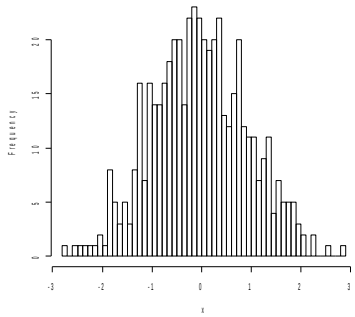


diagramme en bâtons !









Histogramme

~

Densité de probabilité

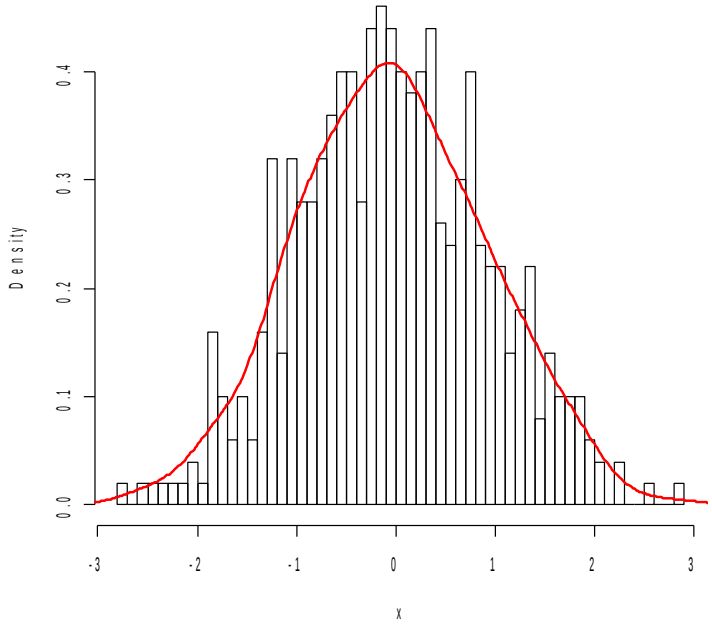
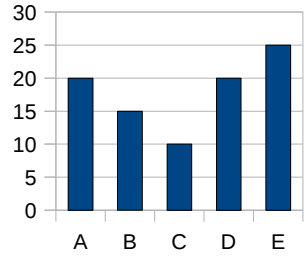
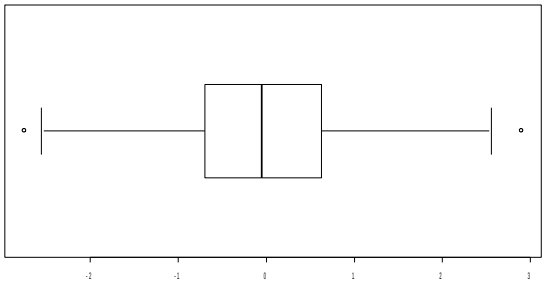
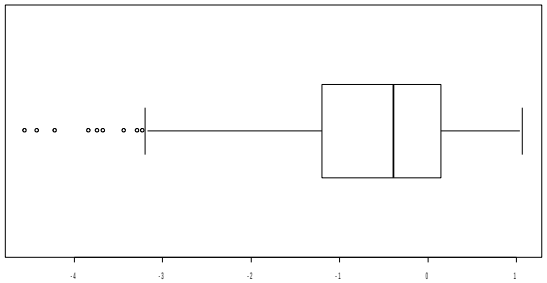
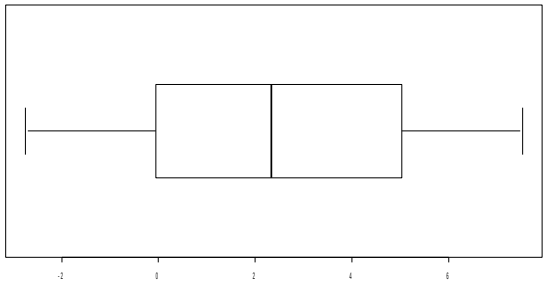
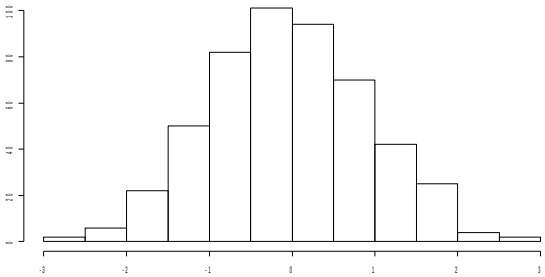
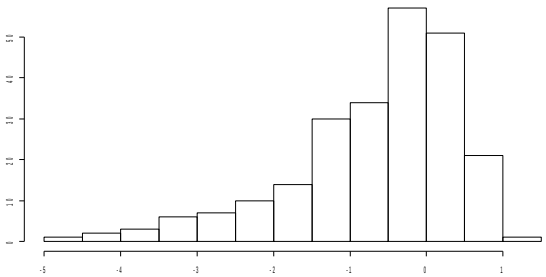
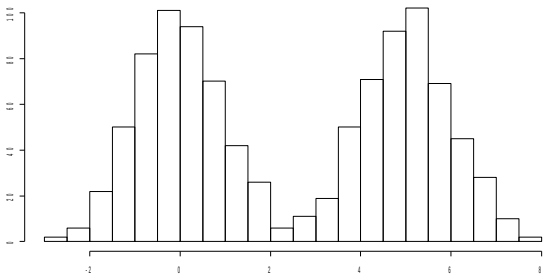


Diagramme en bâtons



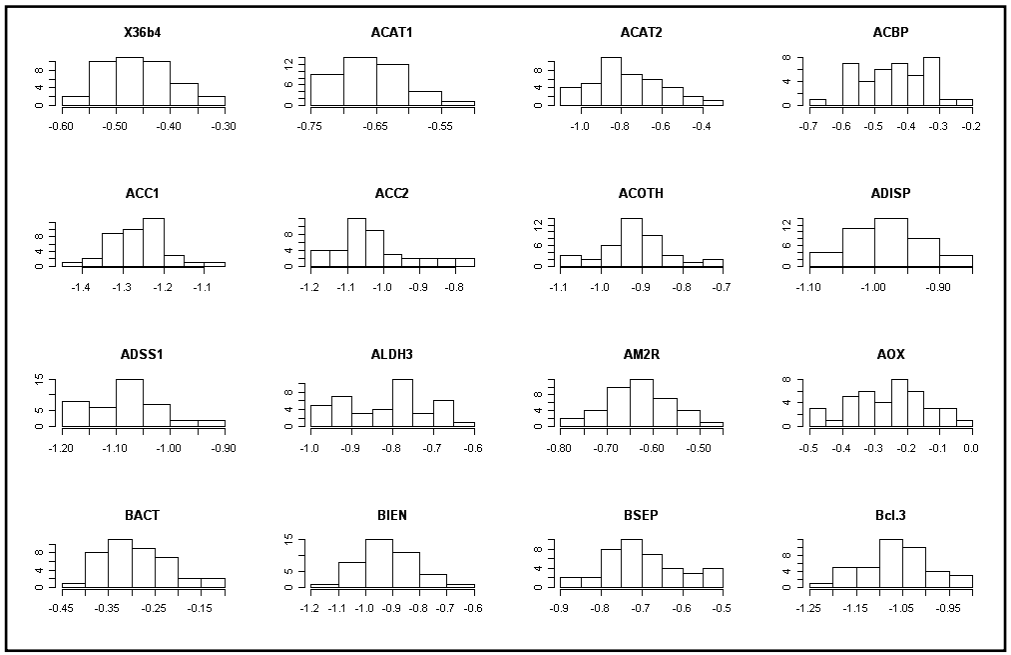
Boxplot et histogramme

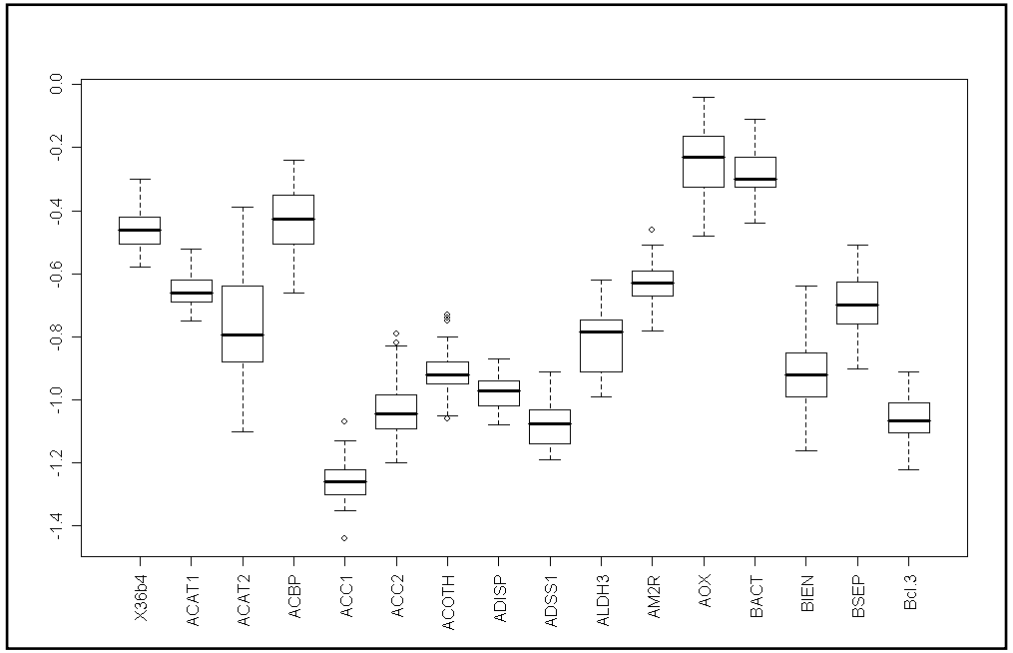







Individuellement, ces graphiques apportent la même information sur la position et la répartition des données. Attention cependant aux cas de multi-modalités que le boxplot ne peut pas capter.

⇒ ***avantage histogramme ?***

Boxplot et histogramme



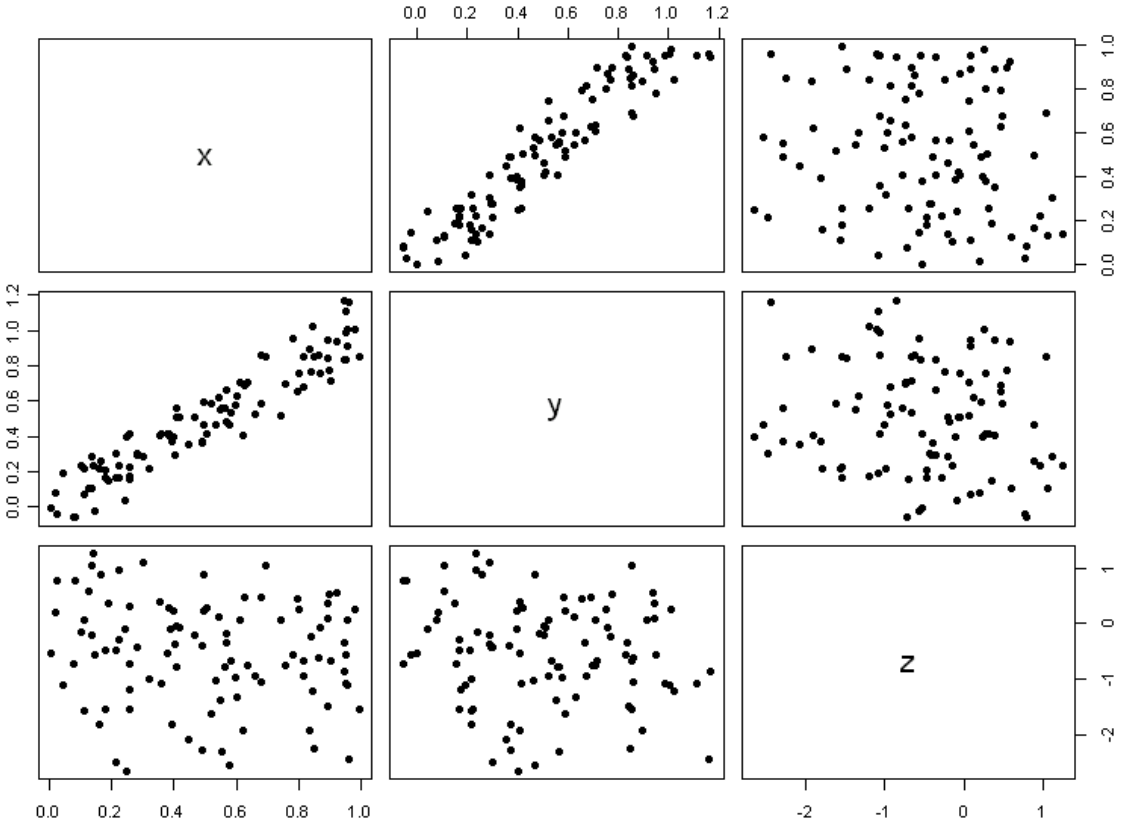


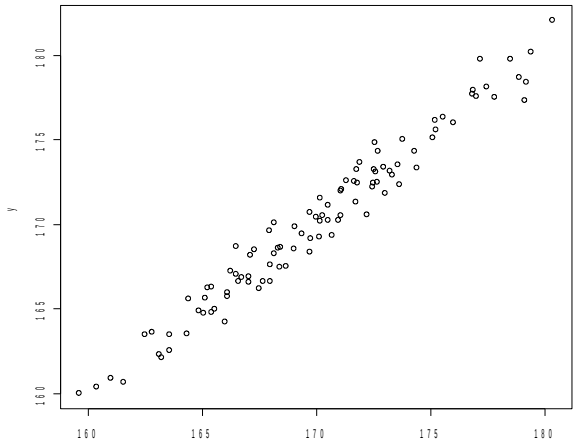
⇒ ***égalité ?***

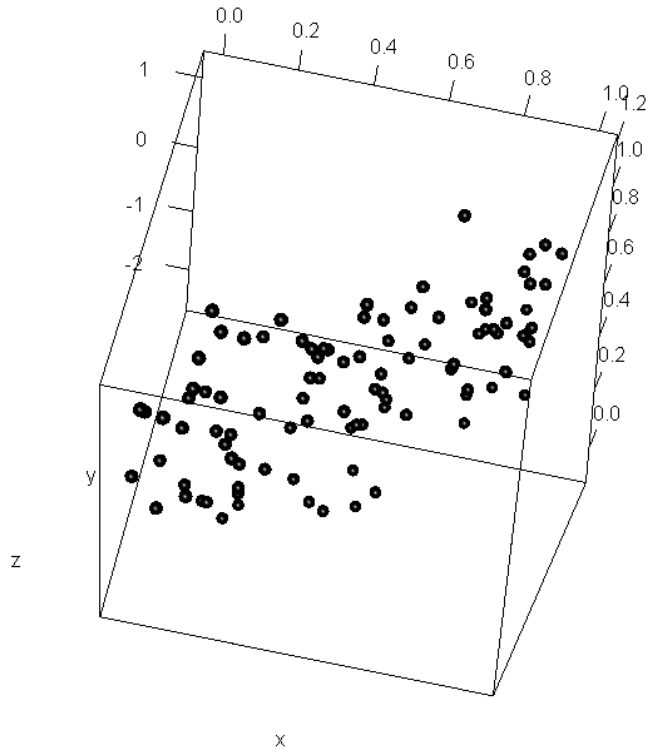
Représentations graphiques

2 variables ou plus

Nuage de points, diagramme de dispersion (*scatterplot*)

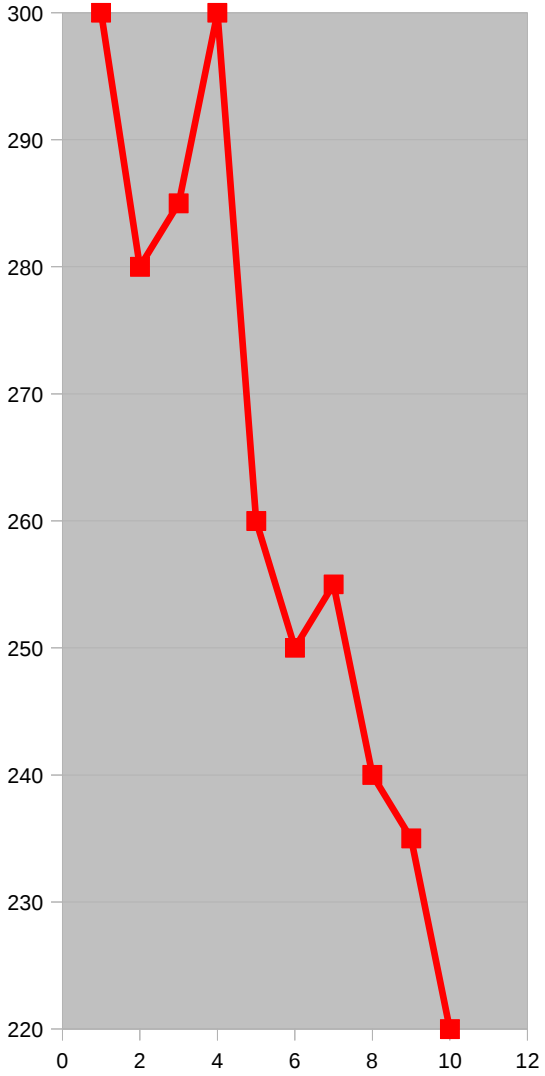


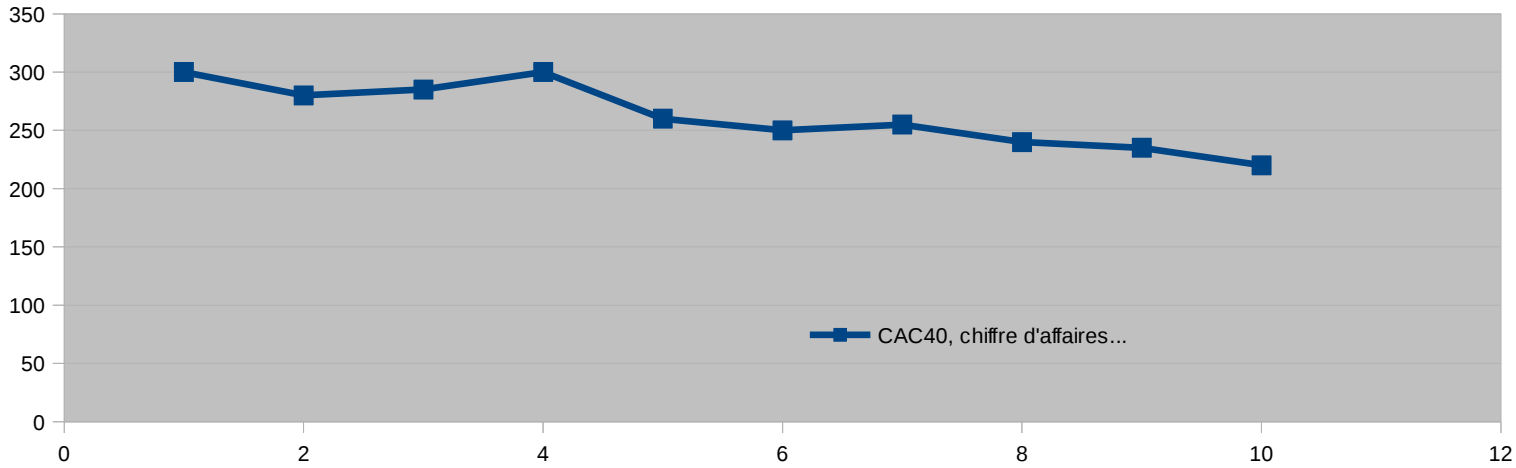




Échelles des axes (1)

Temps	1	2	3	4	5	6	7	8	9	10
Chômage, CAC40, chiffre d'affaires...	300	280	285	300	260	255	255	240	235	220

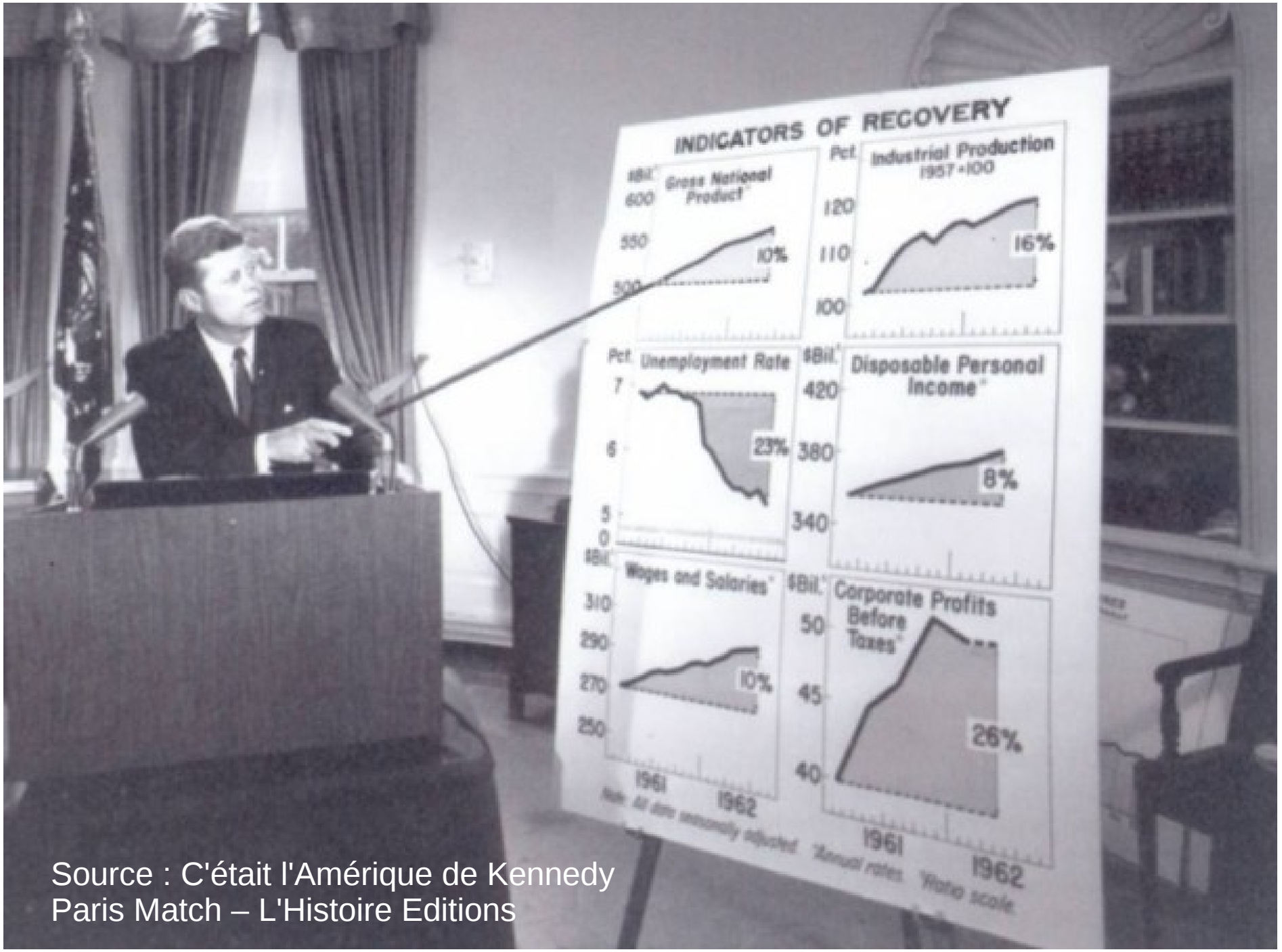




Pas de panique, léger fléchissement...

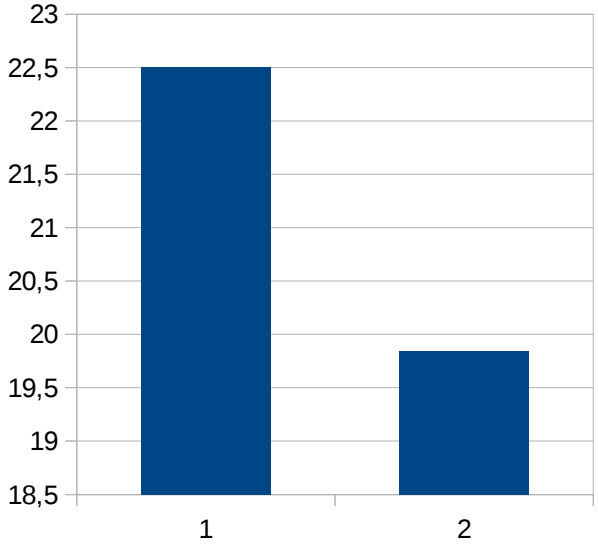
C'est la crise !!! effondrement...

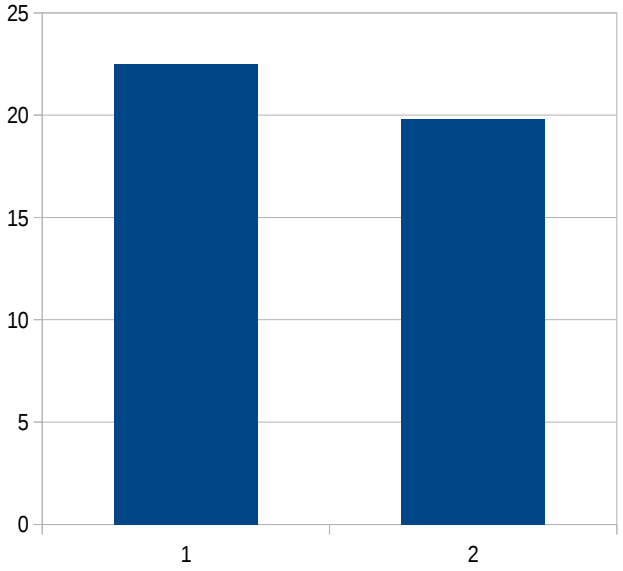
Échelle des axes

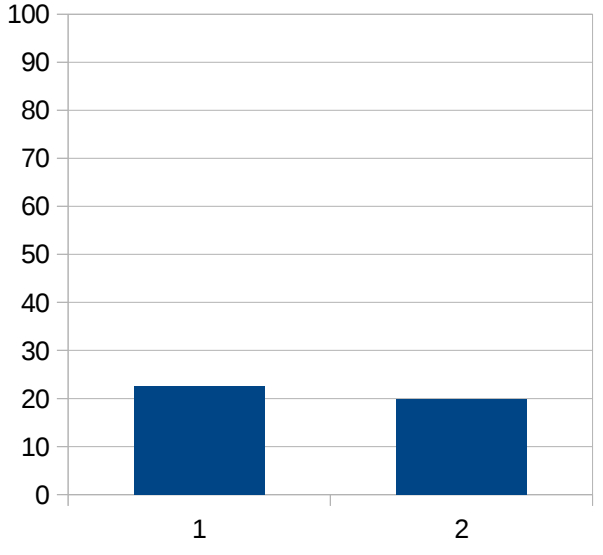


Source : C'était l'Amérique de Kennedy
Paris Match – L'Histoire Editions

Échelle des axes







Conclusion

en.wikipedia.org/wiki/Blind_men_and_an_elephant