

---

# Statistiques

---

---

# Plan

Introduction

Chapitre 1 : Tableaux et méthodes graphiques

Chapitre 2 : Méthodes numériques permettant  
de résumer une série

Chapitre 3 : Indice et taux de croissance

Chapitre 4 : Corrélation et tests de liaison

Chapitre 5 : Régression

---

# bibliographie

B. PY (2007), *La statistique sans formule mathématique*, Pearson Education, 2007

D. ANDERSON, D. SWEENEY et T. WILLIAMS,  
*Statistiques pour l'économie et la gestion*, De Boeck,  
2001

E. BRESSOUD et J.C. KAHANE, *Statistique descriptive avec Excel et la calculatrice*, Pearson Education, 2008

---

# Introduction

---

Qu'est ce que la statistique ?

---

# Exemples de statistiques

L'indice des prix à la consommation a augmenté de 3% sur un an  
(*Source INSEE*)

Le salaire net annuel moyen en France, en 2005, était de 24 446€ pour les hommes et de 19 818€ pour les femmes (*Source INSEE*)

Au 1<sup>er</sup> janvier 2007, les personnes de 20 à 64 ans représentent 58,8% de la population française (*Source INSEE*)

Le taux d'occupation des TGV est de 75% en moyenne en 2007  
(*source SNCF*)

---

# Définition

La statistique c'est l'art et la science de collecter, d'analyser, de présenter et d'interpréter des données

⇒ La statistique permet de résumer et d'interpréter une réalité complexe

⇒ Aide à la prise de décision

---

# Définition

Décrit et synthétise la réalité

⇒ Outil de communication

⇒ permet de faire passer un message

Comment ?

- Sous forme de tableaux
- Sous forme de graphiques
- Sous forme numérique : moyennes, indices, taux de croissance...

---

# Difficultés

- Doit être facile à concevoir et à calculer
- Ne permet pas de décrire tous les profils (moyenne)
- Les indicateurs doivent être neutres et facilement interprétables
- L'interprétations des indicateurs est indispensable

---

# Domaines d'utilisation

- Comptabilité vérification des comptes par sondages
- Finance : comparer plusieurs informations permet la prise de décisions
- Marketing : connaissance des comportements moyen des consommateurs
- Production : contrôle de la qualité
- Economie : visualiser l'état de l'économie

---

# Sources de données

Collecte des données pour une étude statistique est souvent difficile

A partir de bases de données existantes :

- Fichiers internes aux entreprises : volumes des ventes, nombre de clients, effectifs..
- Fichiers externe : les différents ministères ou entreprises privées qui collectent des données (INSEE, EUROSTAT ...)

Par construction de la base de donnée

- Sondages
  - Exhaustifs (recensement)
  - Par échantillon

---

# Statistique descriptive

Ensemble des méthodes qui permettent de décrire les unités statistiques qui composent une population

Représentation par des tableaux, des graphiques ou des données numériques

⇒ Décrit une situation et permet d'en tirer des enseignements

---

# Inférence statistique

Population souvent trop importante

⇒ Pour réduire le coût de collecte, on utilise un échantillon de la population observée

A partir de l'étude de cet échantillon, possibilité d'estimer les comportements ou caractéristiques pour toute la population (contrôle de la qualité)

---

# Vocabulaire

**Population** : ensemble des éléments considérés dans une étude particulière

**Echantillon** : sous-ensemble de la population

**Unité statistique** = élément de la population (individus, animaux, pays...)

La population ou échantillon est décrite selon différents **critères** (données quantitatives) ou **caractères** (données qualitatives).

Chaque caractère peut présenter différentes **modalités** (hommes-femmes pour le sexe, chômeur ou salarié pour le statut...)

Découpage de la population en sous-populations selon différentes **caractéristiques** (âge, sexe, monnaie, superficie...)

# Exemple 1

Données macroéconomiques pour les pays de l'UE à 27 et certains de leurs partenaires commerciaux

	Emissions de gaz à effet de serre en 2003 (en millions de teq CO2)	PIB en 2003 (Milliards d'euros)	Superficie (km2)	Population (en millions)	Population urbaine (en %)	Monnaie
Allemagne (1)	1 030,1	2163,8	357021	82,3	75	euro
Autriche	93,3	223,3023	83858	8,3	67	euro
Belgique	146,3	274,726	30528	10,6	97	euro
Bulgarie	71,2	17,7668	110910	7,7	71	Lev
Chypre	9,3	11,785	9250	1,0	62	euro
Danemark	73,8	188,5003	43094	5,5	72	Couronne danoise
Espagne	410,1	782,929	504762	45,3	77	euro
Estonie	19,7	8,6926	45225	1,3	69	Couronne estonienne
Finlande	84,8	145,938	337030	5,3	62	euro
France	551,9	1594,814	643427	63,6	77	euro
Grèce	133,5	171,4098	131940	11,2	59	euro
Hongrie	80,6	74,5796	93030	10,1	65	Florint
Irlande	68,6	139,4419	70263	4,4	60	euro
Italie	574,1	1335,3537	301320	59,3	68	euro
Lettonie	10,8	9,9778	64569	2,3	68	Lat
Lituanie	21,0	16,4971	35200	3,4	67	Litas
Luxembourg	11,7	25,8343	2585	0,5	83	euro
Malte	3,1	4,4214	315	0,4	95	euro
Pays-Bas	216,3	476,945	41526	16,4	65	euro
Pologne	384,6	191,6438	82931	38,1	62	Zloti
Portugal	83,0	138,5821	312665	10,7	55	euro
République tchèque	145,5	80,9241	78809	10,3	74	Couronne tchèque
Roumanie	156,9	52,613	238391	21,6	55	Leu
Royaume-Uni	658,9	1647,0556	244820	61,0	90	Livre sterling
Slovaquie	50,2	29,4856	48845	5,4	56	Couronne slovaque
Slovénie	19,8	25,7359	20253	2,0	49	euro
Suède	70,7	275,657	449964	9,1	84	Couronne suédoise
<b>Union européenne à 27</b>	<b>5 179,8</b>	<b>10 108,4</b>	<b>4 382 531,0</b>	<b>497,1</b>	-	
Suisse	52,6	287,7538	41290	7,5	68	Franc suisse
Etats-Unis	6 893,8	9689,5332	9826830	302,2	79	Dollar
Japon	1 339,1	3743,5596	377835	127,7	79	Yen
<b>Total de l'échantillon</b>	<b>13 465,4</b>	<b>23 829,3</b>	<b>14 628 486,0</b>	<b>934,5</b>	-	

(1) : incluant l'ex-RDA à partir de 1991.

Source : EUROSTAT et INSEE

---

# Exemple 1

Population = 30 pays ou 30 unités statistiques

Cette population est décrite par 6 critères

# Exemple 2 : tableau croisé

Étudiants des universités par discipline et par cursus (année 2007-2008)

	<b>Cursus Licence</b>	<b>Cursus Master</b>	<b>Cursus Doctorat</b>	Effectif total
	<i>Effectif</i>	<i>Effectif</i>	<i>Effectif</i>	
Droit, sciences politiques	106690	64064	8371	179125
Sciences économiques, gestion (hors AES)	75544	56395	4535	136474
Administration économique et sociale (AES)	30962	7067	0	38029
Lettres, sciences du langage, arts	66541	23525	6932	96998
Langues	84027	17060	2746	103833
Sciences humaines et sociales	135396	63463	14759	213618
Pluri-lettres-langues-sciences humaines	2505	3167	28	5700
Sciences fondamentales et applications	77420	65371	15898	158689
Sciences de la nature et de la vie	39322	19547	10873	69742
Sciences et techniques des activités physiques et sportives	25501	6135	516	32152
Pluri-sciences	20769	1387	145	22301
Médecine - Odontologie	55459	102508	1028	158995
Pharmacie	11752	19560	559	31871
<b>Total hors IUT</b>	<b>731888</b>	<b>449249</b>	<b>66390</b>	<b>1247527</b>
Instituts universitaires de technologie	116223	-	-	116223
<b>Total avec IUT</b>	<b>848111</b>	<b>449249</b>	<b>66390</b>	<b>1363750</b>

Source : INSEE d'après direction de l'Évaluation, de la Prospective et de la Performance (Depp).

---

## Exemple 2 : tableau croisé

Population : étudiants français inscrits à l'université en 2007-2008 (1 363 750 individus)

Représenter selon deux caractères :

- Discipline
- Niveau du cursus

Chaque caractère contient plusieurs modalités

---

# Données quantitatives vs qualitatives

**Données quantitatives** : caractère dénombrables, représentées par des chiffres.

Exemples : superficie, PIB, ventes, CA...

**Données qualitatives** : noms ou étiquettes

Exemples : Monnaie, discipline, cursus

*Remarque : des données numériques peuvent être des données qualitatives*

Exemples : numéro de sécurité sociale, immatriculation, codification numérique des variables ou échelle de valeur (bon = 3, moyen = 2, mauvais = 0)

Distinction importante car toutes les opérations arithmétiques ne sont pas possibles avec des variables qualitatives

---

---

# Variables discrètes et variables continues

**Variables discrètes** : modalités ne peuvent prendre que certaines valeurs

**Variables continues** : variable peut prendre n'importe quelle valeur

Exemples : cursus, nombre d'enfants = variable discrète  
Superficie, PIB = variable continue

---

# Données en coupe transversale et données en séries temporelles

Données en coupe transversale : données collectées à peu près au même moment ou pour une même période (année, mois, jours...)

Exemples : tableau 1 et tableau 2.

Données en séries temporelles : données collectées sur plusieurs périodes (années, mois, jours...)

# Données en coupe transversale et données en séries temporelles

## Données en séries temporelles

**France**  
**Emissions de gaz à effet de serre (Teg CO<sub>2</sub>)**  
**PIB en volume (en milliards d'euros 2000)**

	<b>2000</b>	<b>2001</b>	<b>2002</b>	<b>2003</b>	<b>2004</b>	<b>2005</b>	<b>2006</b>
<b>Emissions</b>	555,6	557,6	548,7	551,9	552,3	555,1	541,3
<b>PIB</b>	1441,37	1468,10	1483,18	1499,31	1536,35	1565,48	1599,46

Source : EUROSTAT

---

# Synthèse à partir de l'exemple 1

## Lecture du tableau

- signification des colonnes
- Les total des colonnes a-t-il toujours un sens ?

Quelles informations peut-on extraire de ce tableau ?

## Questions simples

Pourquoi choisir seulement ces pays ?

Quel pays a la plus grande superficie ou la plus grande population ?

Combien de pays utilisent l'euro dans la population ?

---

---

# Synthèse à partir de l'exemple 1

Possibilité de réaliser des regroupements.

Au sein de l'UE

- Population totale qui utilise l'euro ?
- Quel est le revenu total de l'UE ?
- Quelles sont les émissions total de l'UE ?
- Quelles sont les contributions de chaque pays à chaque critère ?
- Revenu moyen et émissions moyennes ? Existe-t-il de grandes disparités ?

Comparaison entre zone euro et hors zone euro

- Quel est le PIB ou les émissions de la zone euro et hors zone euro?
- Même variables en moyennes ?

# Synthèse à partir de l'exemple 1 : contributions

**Contributions de chaque pays de l'UE à 27 (en pourcentage)**

	Emissions de gaz à effet de serre en 2003 (en millions de teq CO2)	PIB en 2003 (Milliards d'euros)	Superficie (km2)	Population (en millions)
Allemagne (1)	19,89	21,41	8,15	16,56
Autriche	1,80	2,21	1,91	1,67
Belgique	2,82	2,72	0,70	2,13
Bulgarie	1,38	0,18	2,53	1,55
Chypre	0,18	0,12	0,21	0,20
Danemark	1,42	1,86	0,98	1,11
Espagne	7,92	7,75	11,52	9,11
Estonie	0,38	0,09	1,03	0,26
Finlande	1,64	1,44	7,69	1,07
France	10,65	15,78	14,68	12,79
Grèce	2,58	1,70	3,01	2,25
Hongrie	1,56	0,74	2,12	2,03
Irlande	1,33	1,38	1,60	0,89
Italie	11,08	13,21	6,88	11,93
Lettonie	0,21	0,10	1,47	0,46
Lituanie	0,41	0,16	0,80	0,68
Luxembourg	0,23	0,26	0,06	0,10
Malte	0,06	0,04	0,01	0,08
Pays-Bas	4,18	4,72	0,95	3,30
Pologne	7,42	1,90	1,89	7,66
Portugal	1,60	1,37	7,13	2,15
République tchèque	2,81	0,80	1,80	2,07
Roumanie	3,03	0,52	5,44	4,35
Royaume-Uni	12,72	16,29	5,59	12,27
Slovaquie	0,97	0,29	1,11	1,09
Slovénie	0,38	0,25	0,46	0,40
Suède	1,37	2,73	10,27	1,83
<b>Union européenne à 27</b>	<b>100,00</b>	<b>100,00</b>	<b>100,00</b>	<b>100,00</b>

(1) : incluant l'ex-RDA à partir de 1991.

Source : EUROSTAT et INSEE et calculs

# Synthèse à partir de l'exemple 1 : moyennes et dispersions

**Statistiques résumées pour l'UE à 27**

	Emissions de gaz à effet de serre en 2003 (en millions de teq CO2)	PIB en 2003 (Milliards d'euros)	Superficie (km2)	Population (en millions)	Densité moyenne (en hab./km2)	Population urbaine (en %)
Moyenne	191,85	324,39	162315,96	3,05	113	-
médiane	83 (Portugal)	139,4 (Irlande)	312665 (Pologne)	8,3 (Suède)	99 (Autriche)	68 (Lettonie)
valeur maximale	1030,1 (Allemagne)	2163,8 (Allemagne)	643527 (France)	82,3 (Allemagne)	1270 (Malte)	97 (Belgique)
valeur minimale	3,1 (Malte)	4,4 (Malte)	315 (Malte)	0,4 (Malte)	16 (Finlande)	49 (Slovénie)

# Synthèse à partir de l'exemple 1 : dispersions

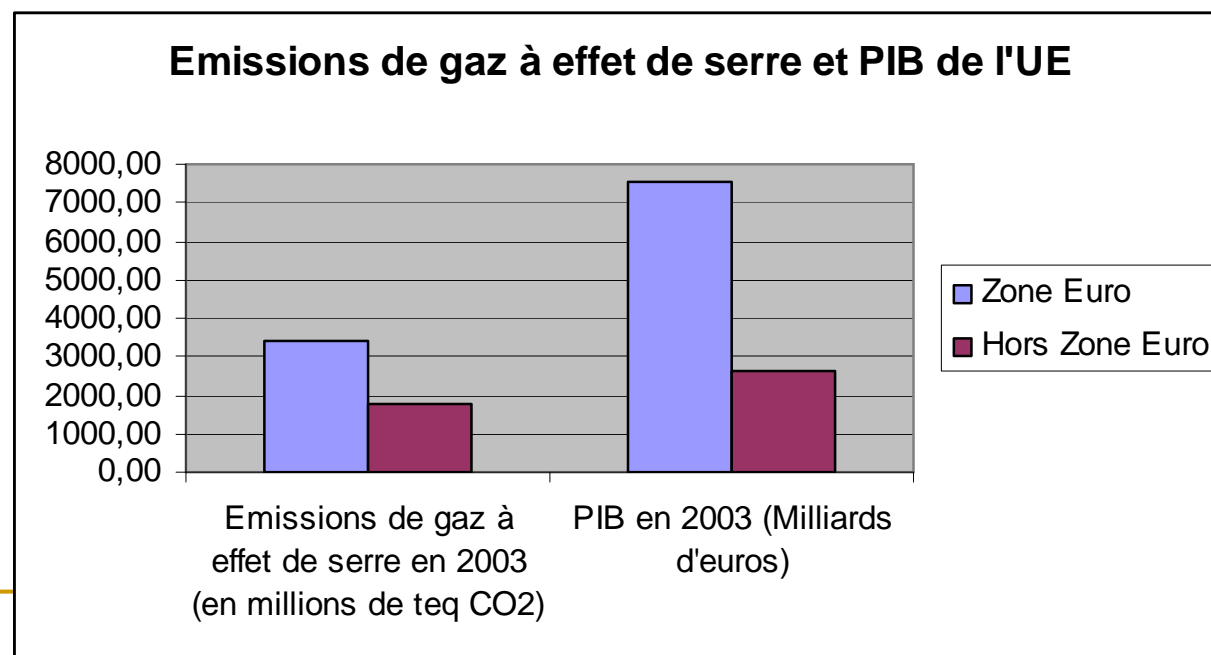
**Déciles de PIB et de PIB par habitant**

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
PIB	9,98 <i>Lettonie</i>	17,77 <i>Bulgarie</i>	29,49 <i>Slovaquie</i>	80,92 <i>République tchèque</i>	145,94 <i>Finlande</i>	191,64 <i>Pologne</i>	275,66 <i>Suède</i>	782,93 <i>Espagne</i>	1647,06 <i>Royaume-Uni</i>	9689,53 <i>Etats-Unis</i>
PIB/habitant	4,34 <i>Lettonie</i>	5,46 <i>Slovaquie</i>	7,86 <i>République tchèque</i>	12,87 <i>Slovénie</i>	17,28 <i>Espagne</i>	25,92 <i>Belgique</i>	27,00 <i>Royaume-Uni</i>	29,32 <i>Japon</i>	32,06 <i>Etats-Unis</i>	51,67 <i>Luxembourg</i>

# Synthèse à partir de l'exemple 1 : regroupements

## Regroupements par zone au sein de l'UE

		Emissions de gaz à effet de serre en 2003 (en millions de teq CO2)	PIB en 2003 (Milliards d'euros)	Superficie (km2)	Population (en millions)
Europe	Zone Euro	3435,87	7515,02	2846743	321,30
	Hors Zone Euro	1743,96	2593,39	1535788	175,80
Total		5179,83	10108,41	4382531	497,10



---

# Synthèse à partir de l'exemple 1

Questions nécessitant des informations complémentaires

- Qui est le plus riche ou qui produit le plus ?
- Qui pollue le plus ?

Ces informations sont-elles pertinentes ? Il faut les interpréter

En terme de production, comparez

- Pologne et Danemark
- Slovénie et Luxembourg

En terme de pollution, comparez

- Danemark et Slovaquie
- Belgique et république Tchèque

# Synthèse à partir de l'exemple 1

Données macroéconomiques pour les pays de l'UE à 27 et certains de leurs partenaires commerciaux

	Emissions de gaz à effet de serre en 2003 (en millions de teq CO2)	PIB en 2003 (Milliards d'euros)	Superficie (km2)	Population (en millions)	Densité moyenne (en hab./km2)	Population urbaine (en %)	PIB/habitant (en milliers d'euros)	Pollution par habitant (en Teq CO2)	pollution/PIB (en kg eq CO2 par euro)	Monnaie
Allemagne (1)	1 030,1	2163,8	357021	82,3	231	75	26,29	12,52	0,48	euro
Autriche	93,3	223,3023	83858	8,3	99	67	26,90	11,24	0,42	euro
Belgique	146,3	274,726	30528	10,6	347	97	25,92	13,80	0,53	euro
Bulgarie	71,2	17,7668	110910	7,7	69	71	2,31	9,25	4,01	Lev
Chypre	9,3	11,785	9250	1,0	108	62	11,79	9,30	0,79	euro
Danemark	73,8	188,5003	43094	5,5	128	72	34,27	13,41	0,39	Couronne danoise
Espagne	410,1	782,929	504762	45,3	90	77	17,28	9,05	0,52	euro
Estonie	19,7	8,6926	45225	1,3	29	69	6,69	15,15	2,27	Couronne estonienne
Etats-Unis	6 893,8	9689,5332	9826830	302,2	31	79	32,06	22,81	0,71	Dollar
Finlande	84,8	145,938	337030	5,3	16	62	27,54	16,00	0,58	euro
France	551,9	1594,814	643427	63,6	99	77	25,08	8,68	0,35	euro
Grèce	133,5	171,4098	131940	11,2	85	59	15,30	11,92	0,78	euro
Hongrie	80,6	74,5796	93030	10,1	109	65	7,38	7,98	1,08	Florint
Irlande	68,6	139,4419	70263	4,4	63	60	31,69	15,60	0,49	euro
Italie	574,1	1335,3537	301320	59,3	197	68	22,52	9,68	0,43	euro
Japon	1 339,1	3743,5596	377835	127,7	338	79	29,32	10,49	0,36	Yen
Lettonie	10,8	9,9778	64569	2,3	36	68	4,34	4,72	1,09	Lat
Lituanie	21,0	16,4971	35200	3,4	97	67	4,85	6,18	1,27	Litas
Luxembourg	11,7	25,8343	2585	0,5	193	83	51,67	23,33	0,45	euro
Malte	3,1	4,4214	315	0,4	1 270	95	11,05	7,65	0,69	euro
Pays-Bas	216,3	476,945	41526	16,4	395	65	29,08	13,19	0,45	euro
Pologne	384,6	191,6438	82931	38,1	459	62	5,03	10,09	2,01	Zloti
Portugal	83,0	138,5821	312665	10,7	34	55	12,95	7,76	0,60	euro
République tchèque	145,5	80,9241	78809	10,3	131	74	7,86	14,13	1,80	Couronne tchèque
Roumanie	156,9	52,613	238391	21,6	91	55	2,44	7,26	2,98	Leu
Royaume-Uni	658,9	1647,0556	244820	61,0	249	90	27,00	10,80	0,40	Livre sterling
Slovaquie	50,2	29,4856	48845	5,4	111	56	5,46	9,30	1,70	Couronne slovaque
Slovénie	19,8	25,7359	20253	2,0	99	49	12,87	9,89	0,77	euro
Suède	70,7	275,657	449964	9,1	20	84	30,29	7,77	0,26	Couronne suédoise
Suisse	52,6	287,7538	41290	7,5	182	68	38,37	7,02	0,18	Franc suisse
<b>Union européenne à 27</b>	<b>12 195,7</b>	<b>21 167,4</b>	<b>14 157 079,0</b>	<b>833,3</b>	<b>59</b>	<b>-</b>	<b>25,40</b>	<b>14,64</b>	<b>0,58</b>	

(1) : incluant l'ex-RDA à partir de 1991.

Source : EUROSTAT et INSEE

---

# Synthèse à partir de l'exemple 1

Existe-t-il des liaisons statistiques permettant d'expliquer des résultats?

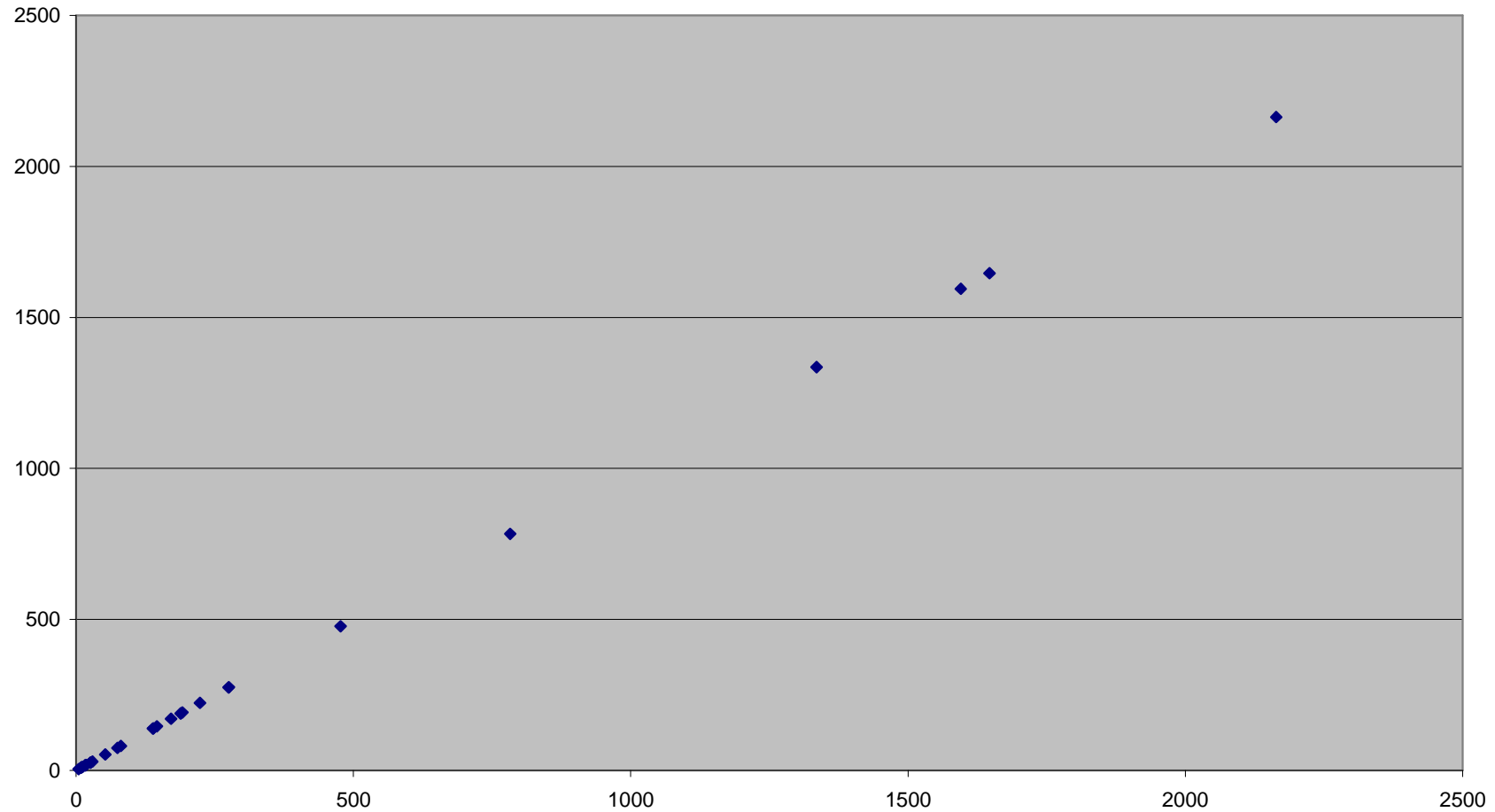
Lien entre population et PIB ?

Lien entre pollution et PIB ?

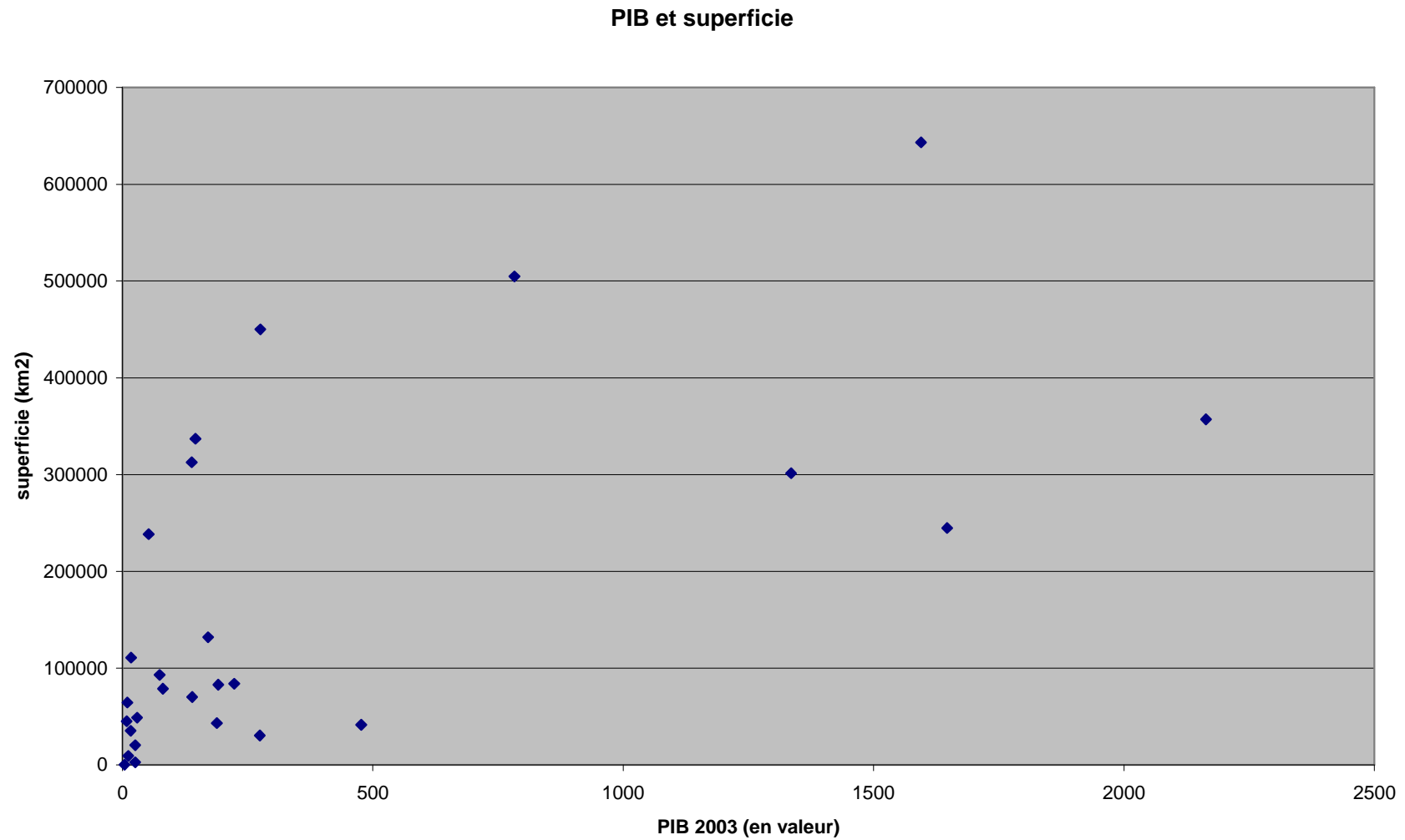
Lien entre pollution et densité de pollution ?

# Synthèse à partir de l'exemple 1 : liaison

Exemple de liaison parfaite

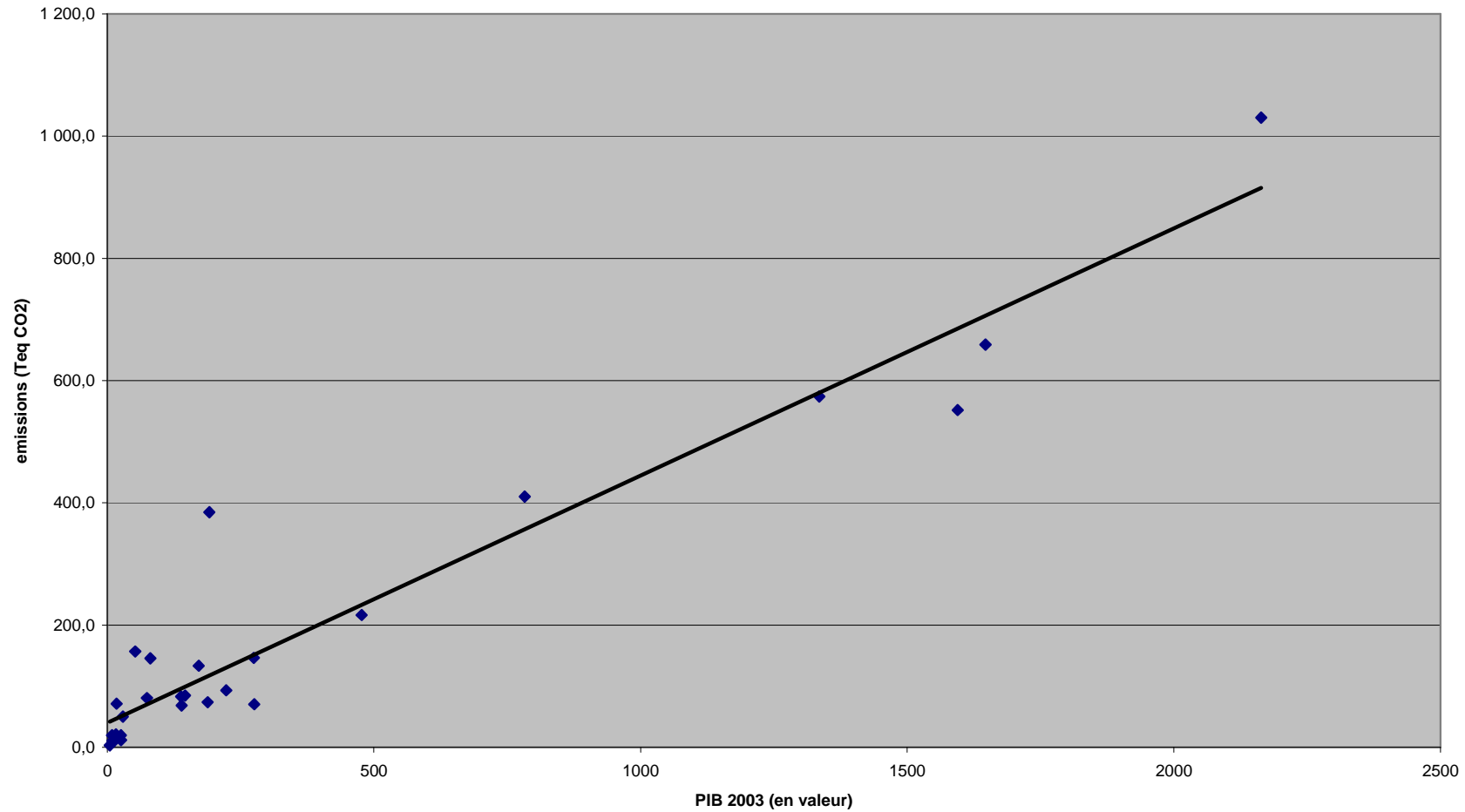


# Synthèse à partir de l'exemple 1 : liaison



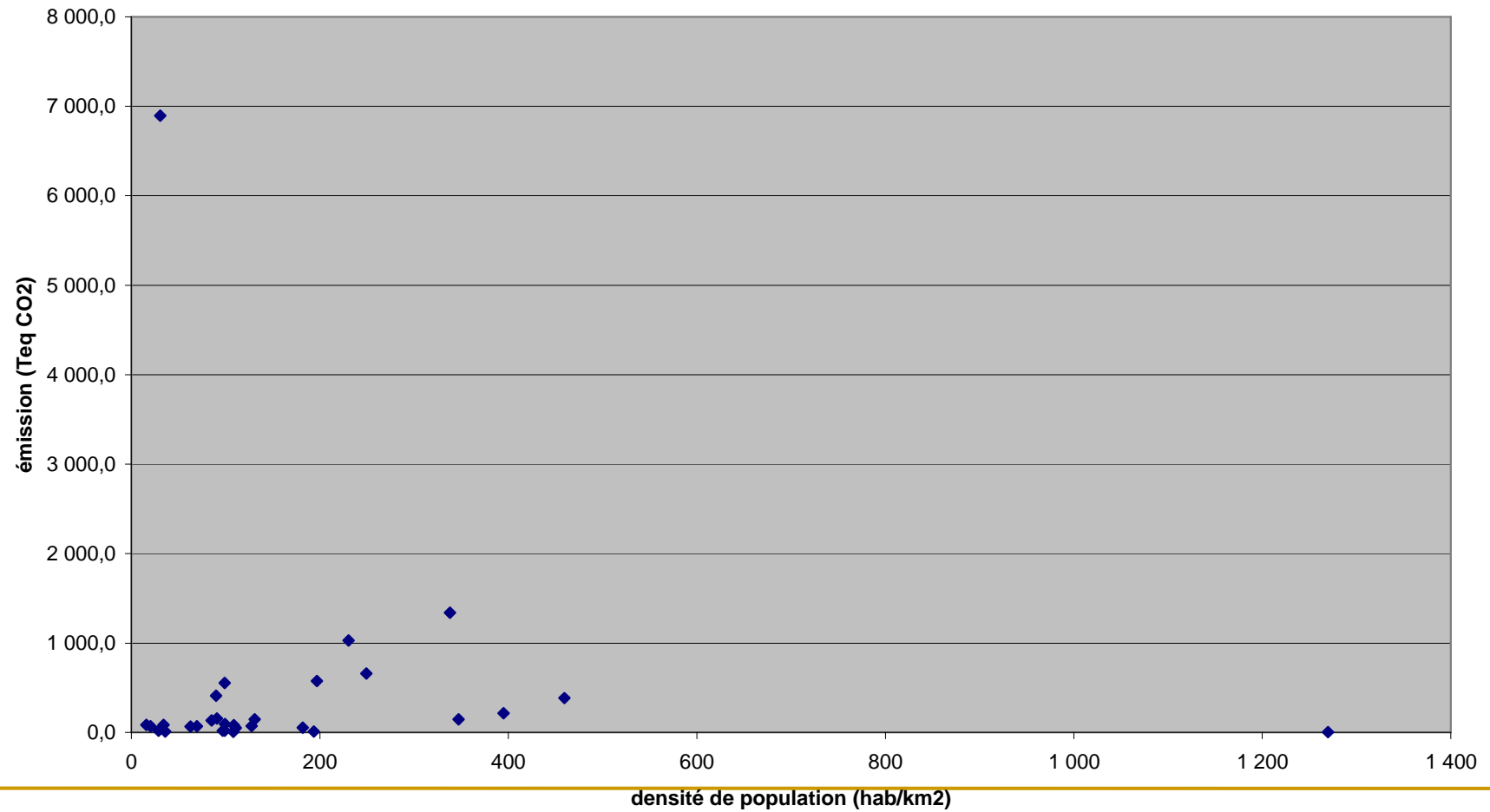
# Synthèse à partir de l'exemple 1 : liaison

Emissions et PIB pour chaque pays



# Synthèse à partir de l'exemple 1 : liaison

Emissions et densité de population



---

# Chapitre 1 : tableaux et graphiques

---

---

# Plan

1. Introduction :
  - ❑ Lecture de tableaux
  - ❑ Construction de tableaux et de graphiques
  
- Données qualitatives
  
- Données quantitatives

# Introduction : Lecture d'un tableau

## Étudiants des universités par discipline et par cursus (année 2007-2008)

	<b>Cursus Licence</b>	<b>Cursus Master</b>	<b>Cursus Doctorat</b>	Effectif total
	<i>Effectif</i>	<i>Effectif</i>	<i>Effectif</i>	
Droit, sciences politiques	106690	64064	8371	179125
Sciences économiques, gestion (hors AES)	75544	56395	4535	136474
Administration économique et sociale (AES)	30962	7067	0	38029
Lettres, sciences du langage, arts	66541	23525	6932	96998
Langues	84027	17060	2746	103833
Sciences humaines et sociales	135396	63463	14759	213618
Pluri-lettres-langues-sciences humaines	2505	3167	28	5700
Sciences fondamentales et applications	77420	65371	15898	158689
Sciences de la nature et de la vie	39322	19547	10873	69742
Sciences et techniques des activités physiques et sportives	25501	6135	516	32152
Pluri-sciences	20769	1387	145	22301
Médecine - Odontologie	55459	102508	1028	158995
Pharmacie	11752	19560	559	31871
<b>Total hors IUT</b>	<b>731888</b>	<b>449249</b>	<b>66390</b>	<b>1247527</b>
Instituts universitaires de technologie	116223	-	-	116223
<b>Total avec IUT</b>	<b>848111</b>	<b>449249</b>	<b>66390</b>	<b>1363750</b>

Source : INSEE d'après direction de l'Évaluation, de la Prospective et de la Performance (Depp).

---

# Introduction : Lecture d'un tableau

- Titre et organisation :
  - Quelles sont les données représentées ? Quelles sont les modalités ?
- Source du tableau : la provenance des données est-elle fiable ?
- Contenu du tableau :
  - Quelle est l'unité des variables ?
  - Lecture en ligne et/ou en colonne ?
  - Lecture rapide : chiffres extrêmes...
  - Le travail d'analyse et d'interprétation peut alors commencer

---

# Introduction : Construction d'un tableau

Quatre principes fondamentaux pour la présentation d'un tableau

- Le titre : le plus précis possible
- La source des données
- L'intitulé des lignes et colonnes
- Les unités des variables

---

# Introduction : Construction d'un graphique

Graphique doit être compris très rapidement

- Titre explicite
- Axes explicites : unités et intitulés
- Ne doit pas contenir trop d'informations

## 2. Données qualitatives : tableau unidimensionnel

Données (fictives) d'un échantillon de 50 achats de boisson non alcoolisée

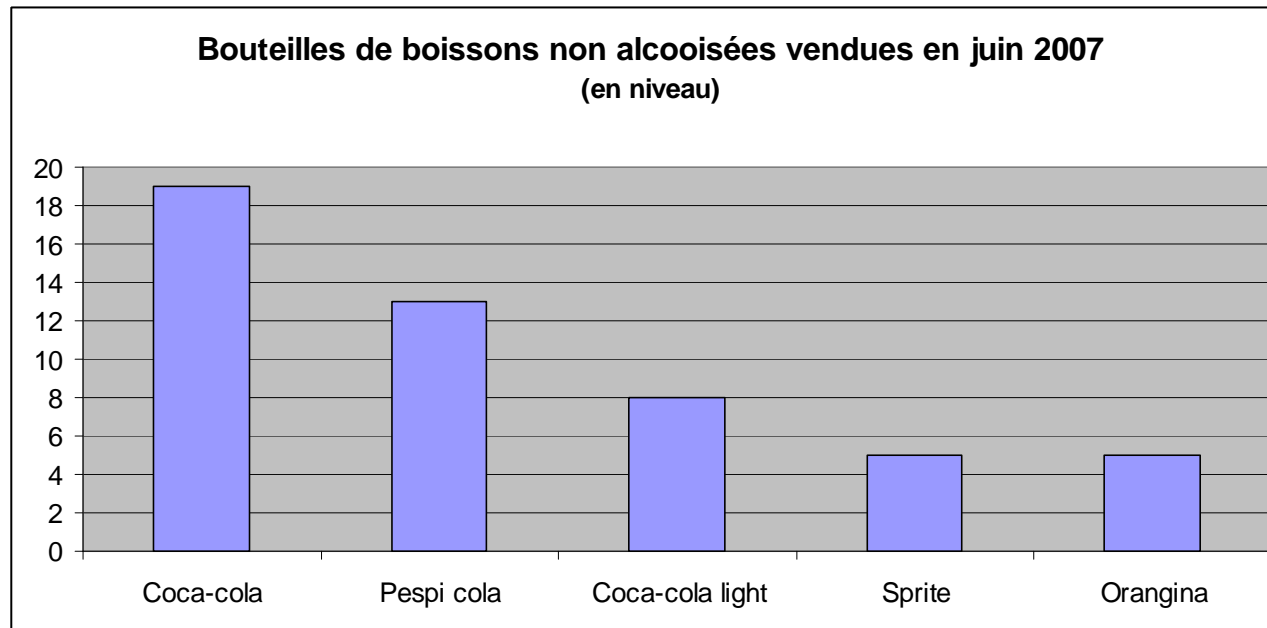
Boisson	nombre de bouteilles vendues	fréquence relative	Fréquence (en %)	Fréquence cumulée
Coca-cola	19	0,38	38	38
Pespi cola	13	0,26	26	64
Coca-cola light	8	0,16	16	80
Sprite	5	0,1	10	90
Orangina	5	0,1	10	100
<b>Effectif total</b>	<b>50</b>	<b>1</b>	<b>100</b>	

source : D. ANDERSON, D. SWEENEY et T. WILLIAMS (2001)

$$\text{Fréquence relative} = \frac{\text{Effectif de la modalité } x}{\text{effectif total}}$$

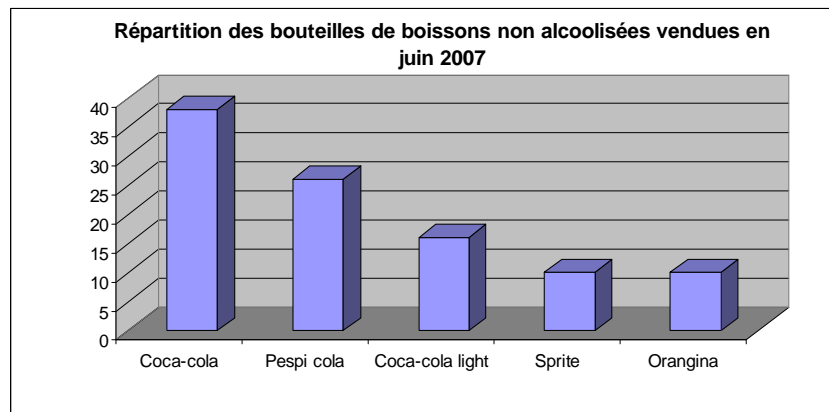
$$\text{Fréquence relative} = \frac{\text{Effectif de la modalité } x}{\text{effectif total}} \times 100$$

## 2. Données qualitatives : graphiques

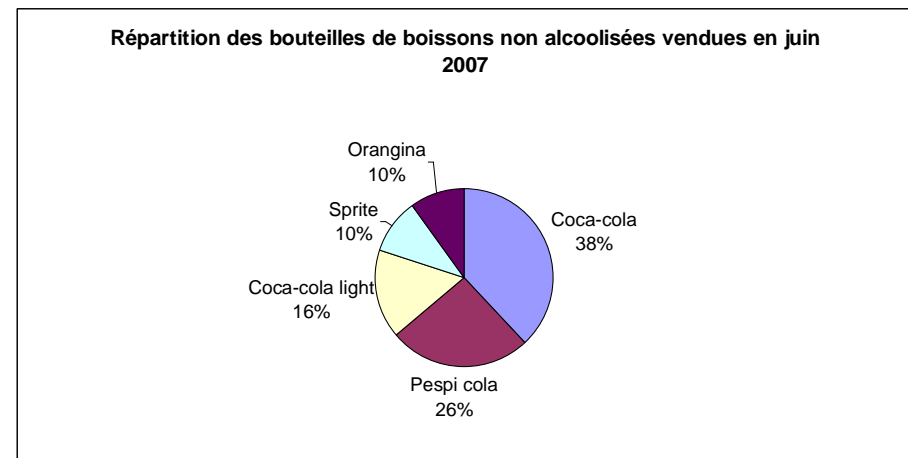


## 2. Données qualitatives : graphiques

Toutes les barres doivent avoir la même largeur et l'espace entre les barres doit être le même. Réduit le risque de mauvaise interprétation



Taille des secteurs : coca représente un angle de  $0,38 \times 360 = 136,8^\circ$



## 2. Données qualitatives : tableaux pluri-dimensionnels

Répartition des étudiants des universités françaises selon la discipline et le cursus  
(Année 2007-2008)

	Cursus Licence	Cursus Master	Cursus Doctorat	Fréquence totale
	Fréquence	Fréquence	Fréquence	
Droit, sciences politiques	7,82	4,70	0,61	<b>13,13</b>
Sciences économiques, gestion (hors AES)	5,54	4,14	0,33	<b>10,01</b>
Administration économique et sociale (AES)	2,27	0,52	0,00	<b>2,79</b>
Lettres, sciences du langage, arts	4,88	1,73	0,51	<b>7,11</b>
Langues	6,16	1,25	0,20	<b>7,61</b>
Sciences humaines et sociales	9,93	4,65	1,08	<b>15,66</b>
Pluri-lettres-langues-sciences humaines	0,18	0,23	0,00	<b>0,42</b>
Sciences fondamentales et applications	5,68	4,79	1,17	<b>11,64</b>
Sciences de la nature et de la vie	2,88	1,43	0,80	<b>5,11</b>
STAPS	1,87	0,45	0,04	<b>2,36</b>
Pluri-sciences	1,52	0,10	0,01	<b>1,64</b>
Médecine - Odontologie	4,07	7,52	0,08	<b>11,66</b>
Pharmacie	0,86	1,43	0,04	<b>2,34</b>
<b>Total hors IUT</b>	<b>53,67</b>	<b>32,94</b>	<b>4,87</b>	<b>91,48</b>
Instituts universitaires de technologie	8,52	//	//	<b>8,52</b>
<b>Total avec IUT</b>	<b>62,19</b>	<b>32,94</b>	<b>4,87</b>	<b>100</b>

/// : absence de résultat due à la nature des choses.

Champ : France.

Source : direction de l'Évaluation, de la Prospective et de la Performance (Depp).

## 2. Données qualitatives : tableaux pluri-dimensionnels

Répartition des étudiants des universités françaises selon la discipline par cursus  
(Année 2007-2008)

	Cursus Licence	Cursus Master	Cursus Doctorat	Fréquence totale
	Fréquence	Fréquence	Fréquence	
Droit, sciences politiques	12,58	14,26	12,61	<b>13,13</b>
Sciences économiques, gestion (hors AES)	8,91	12,55	6,83	<b>10,01</b>
Administration économique et sociale (AES)	3,65	1,57	0,00	<b>2,79</b>
Lettres, sciences du langage, arts	7,85	5,24	10,44	<b>7,11</b>
Langues	9,91	3,80	4,14	<b>7,61</b>
Sciences humaines et sociales	15,96	14,13	22,23	<b>15,66</b>
Pluri-lettres-langues-sciences humaines	0,30	0,70	0,04	<b>0,42</b>
Sciences fondamentales et applications	9,13	14,55	23,95	<b>11,64</b>
Sciences de la nature et de la vie	4,64	4,35	16,38	<b>5,11</b>
STAPS	3,01	1,37	0,78	<b>2,36</b>
Pluri-sciences	2,45	0,31	0,22	<b>1,64</b>
Médecine - Odontologie	6,54	22,82	1,55	<b>11,66</b>
Pharmacie	1,39	4,35	0,84	<b>2,34</b>
<b>Total hors IUT</b>	<b>86,30</b>	<b>100,00</b>	<b>100,00</b>	<b>91,48</b>
Instituts universitaires de technologie	13,70	//	//	<b>8,52</b>
<b>Total avec IUT</b>	<b>100,00</b>	<b>100,00</b>	<b>100,00</b>	<b>100,00</b>

/// : absence de résultat due à la nature des choses.

Champ : France.

Source : direction de l'Évaluation, de la Prospective et de la Performance (Depp).

## 2. Données qualitatives : tableaux pluri-dimensionnels

Répartition des étudiants des universités françaises selon le cursus par discipline  
(Année 2007-2008)

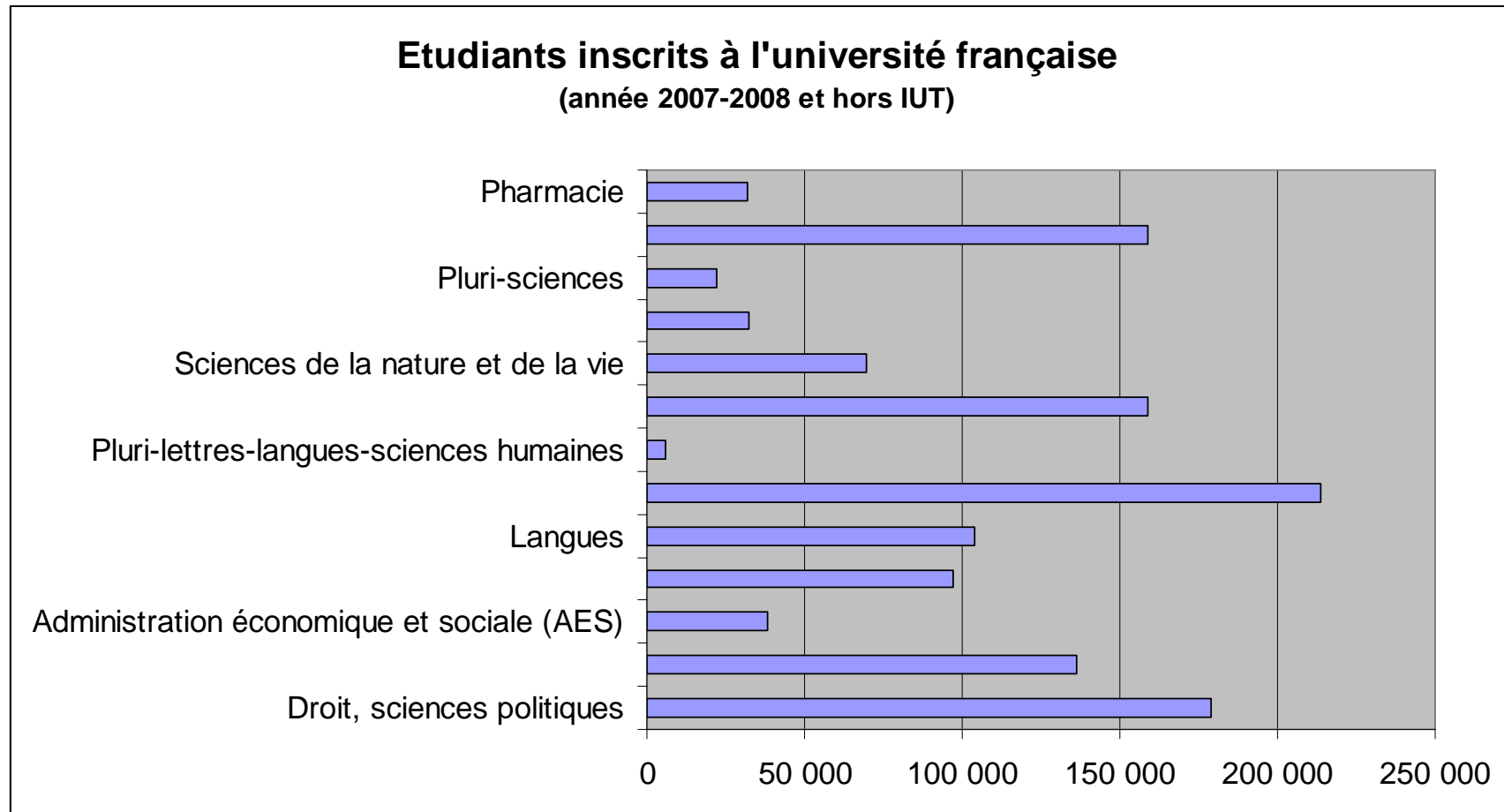
	Cursus Licence	Cursus Master	Cursus Doctorat	Fréquence totale
	Fréquence	Fréquence	Fréquence	
Droit, sciences politiques	59,56	35,76	4,67	100
Sciences économiques, gestion (hors AES)	55,35	41,32	3,32	100
Administration économique et sociale (AES)	81,42	18,58	0,00	100
Lettres, sciences du langage, arts	68,60	24,25	7,15	100
Langues	80,93	16,43	2,64	100
Sciences humaines et sociales	63,38	29,71	6,91	100
Pluri-lettres-langues-sciences humaines	43,95	55,56	0,49	100
Sciences fondamentales et applications	48,79	41,19	10,02	100
Sciences de la nature et de la vie	56,38	28,03	15,59	100
STAPS	79,31	19,08	1,60	100
Pluri-sciences	93,13	6,22	0,65	100
Médecine - Odontologie	34,88	64,47	0,65	100
Pharmacie	36,87	61,37	1,75	100
<b>Total hors IUT</b>	<b>58,67</b>	<b>36,01</b>	<b>5,32</b>	<b>100</b>
Instituts universitaires de technologie	100,00	//	//	100
<b>Total avec IUT</b>	<b>62,19</b>	<b>32,94</b>	<b>4,87</b>	<b>100</b>

/// : absence de résultat due à la nature des choses.

Champ : France.

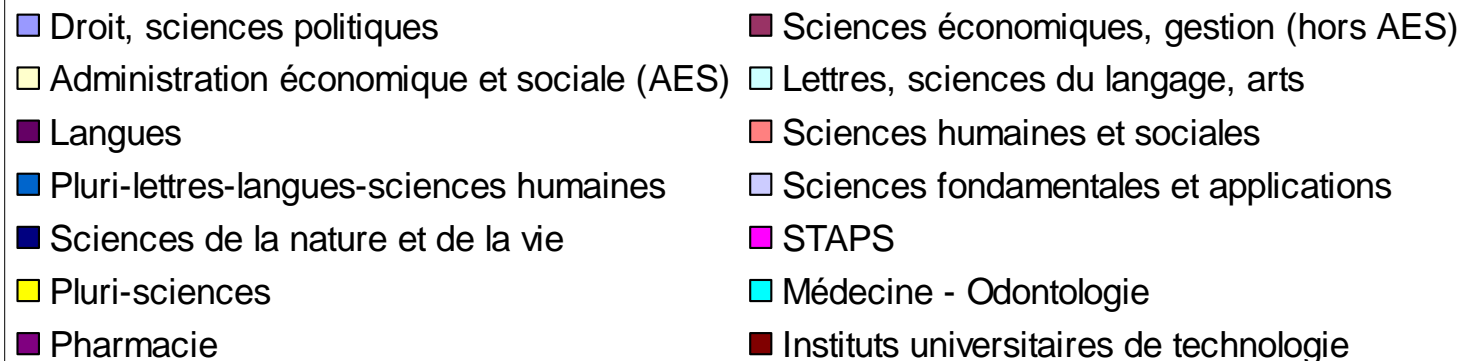
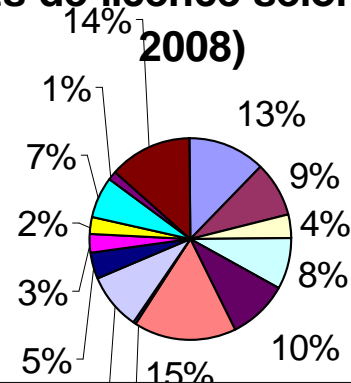
Source : direction de l'Évaluation, de la Prospective et de la Performance (Depp).

## 2. Données qualitatives : graphiques



## 2. Données qualitatives : graphiques

Répartition des étudiants de licence selon la discipline (année 2007-2008)



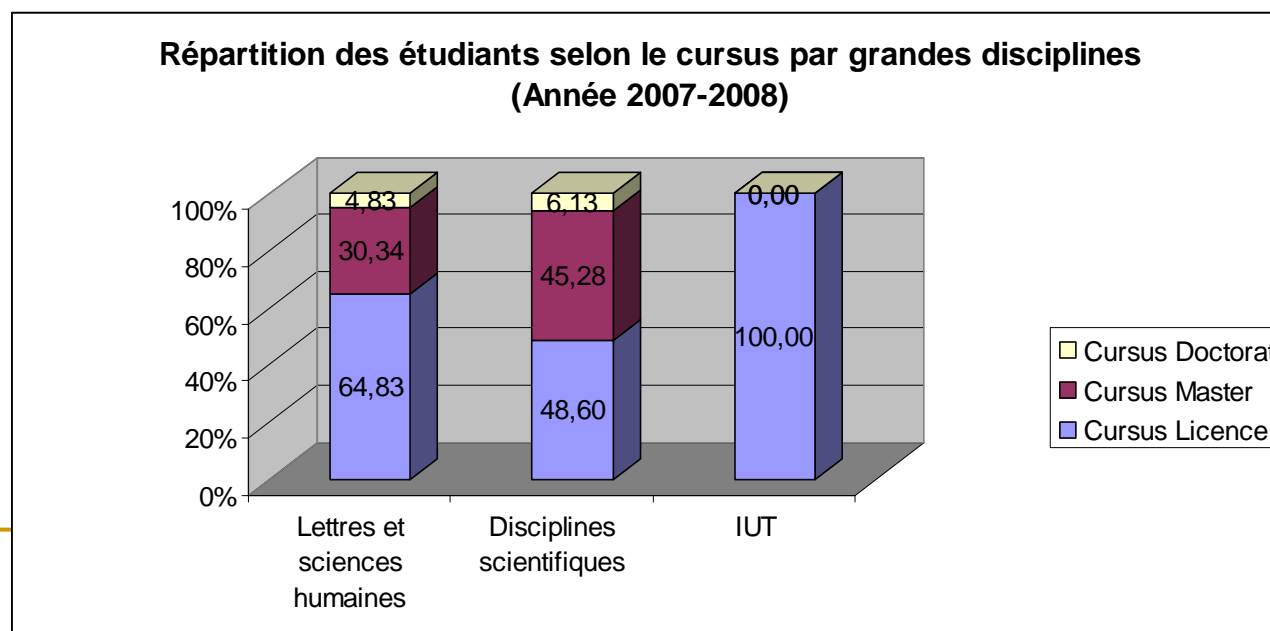
## 2. Données qualitatives : regroupements

Étudiants des universités françaises par discipline en pourcentage (Année 2007-2008)

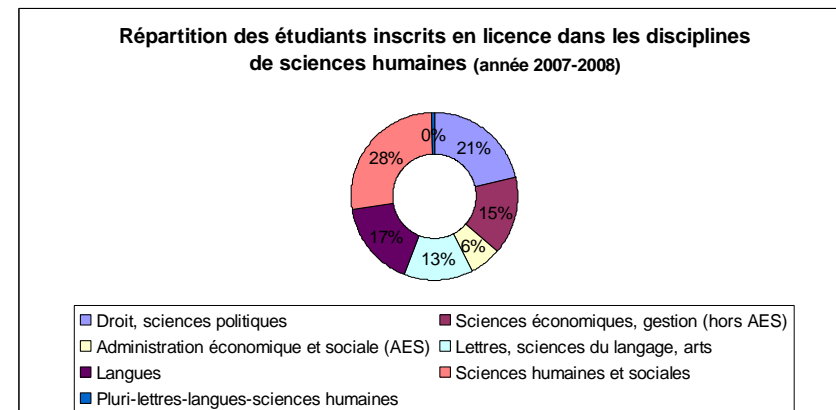
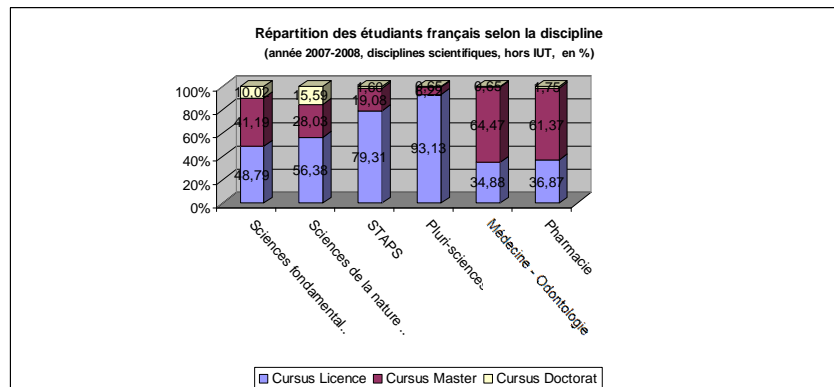
	Cursus Licence	Cursus Master	Cursus Doctorat	Total
Lettres et sciences humaines	64,83	30,34	4,83	100
Disciplines scientifiques	48,60	45,28	6,13	100
IUT	100,00	0,00	0,00	100
<b>Total</b>	62	33	5	100

Champ : France.

Source : direction de l'Évaluation, de la Prospective et de la Performance (Depp).



## 2. Données qualitatives : regroupements



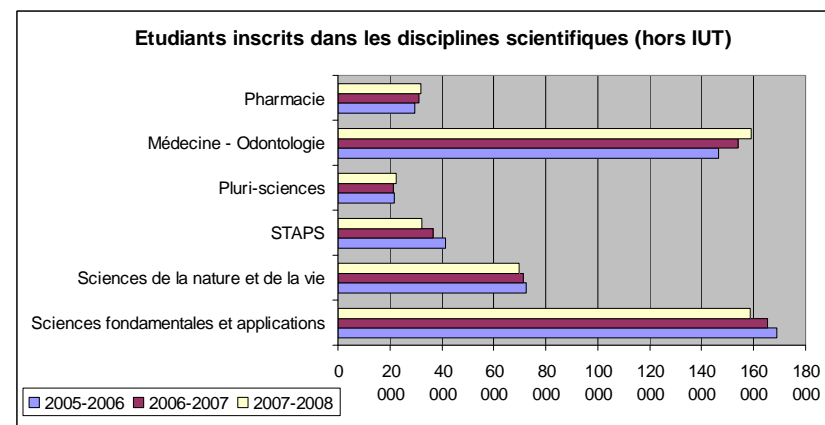
## 2. Données qualitatives : graphiques

Étudiants des universités par discipline

	2005-2006	2006-2007	2007-2008
	Effectif total	Effectif total	Effectif total
Droit, sciences politiques	175 853	178 365	179 125
Sciences économiques, gestion (hors AES)	134 796	134 728	136 474
Administration économique et sociale (AES)	44 451	41 368	38 029
Lettres, sciences du langage, arts	111 452	104 149	96 998
Langues	111 557	108 829	103 833
Sciences humaines et sociales	245 173	232 500	213 618
Pluri-lettres-langues-sciences humaines	4 947	5 576	5 700
Sciences fondamentales et applications	169 158	165 377	158 689
Sciences de la nature et de la vie	72 389	71 320	69 742
STAPS	41 516	36 641	32 152
Pluri-sciences	21 617	21 183	22 301
Médecine - Odontologie	146 589	154 082	158 995
Pharmacie	29 624	31 290	31 871
<b>Total hors IUT</b>	<b>1 309 122</b>	<b>1 285 408</b>	<b>1 247 527</b>
Instituts universitaires de technologie	112 597	113 769	116 223
<b>Total avec IUT</b>	<b>1 421 719</b>	<b>1 399 177</b>	<b>1 363 750</b>

Champ : France.

Source : direction de l'Évaluation, de la Prospective et de la Performance (Depp).



# 3. Données quantitatives

**Durée en jour d'un audit**

12	1
13	1
14	2
15	2
16	1
17	1
18	3
19	1
20	1
21	1
22	2
23	1
27	1
28	1
33	1

source : D. ANDERSON, D. SWEENEY et T. WILLIAMS (2001)

Données trop semblables pour pouvoir les représenter graphiquement

⇒ Regroupements en classes

⇒ Faire ressortir la variation des données

Choix

- Nombre de classes
- Largeur des classes : préférable qu'elles soient de largeurs identiques pour éviter les mauvaises interprétations (pas toujours possible)

# 3. Données quantitatives : regroupements quantitatifs

Choix nombre de classes = 5  $\text{largeur approximative de la classe} = \frac{\text{Valeur la plus élevée} - \text{valeur la plus faible}}{\text{nombre de classes}}$

Chaque donnée ne doit appartenir qu'à une seule et unique classe :

Amplitude de la classe :  $\text{Valeur la plus élevée de la classe} - \text{valeur la plus faible de la classe}$

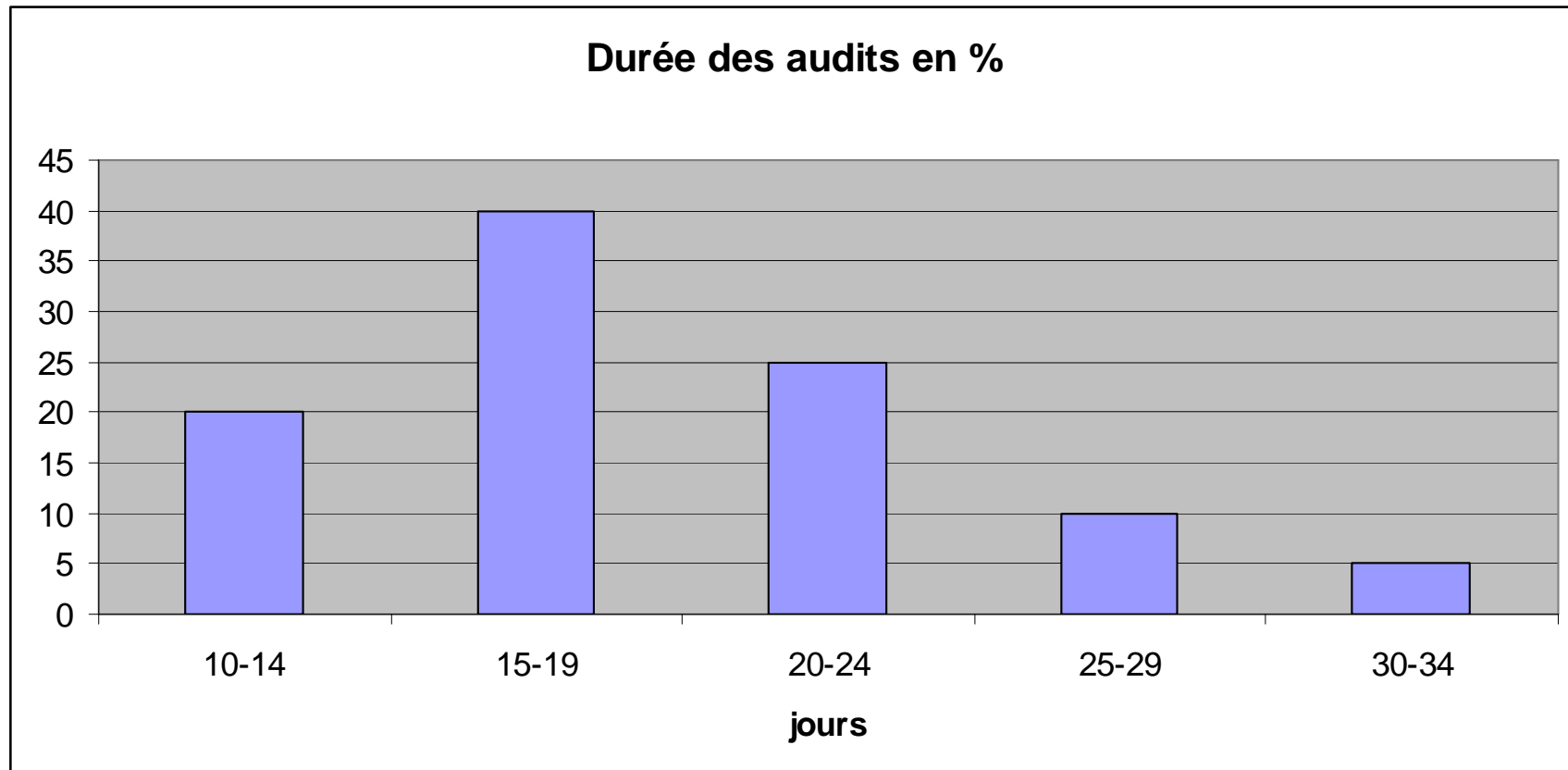
Centre de la classe :  $\text{centre de la classe} = \frac{\text{Valeur la plus élevée} + \text{valeur la plus faible}}{2}$

Distributions pour les données sur les audits

Durée des audits (jours)	Nombre	Fréquence relative	Fréquence en %	Fréquence cumulée
10-14	4	0,2	20	20
15-19	8	0,4	40	60
20-24	5	0,25	25	85
25-29	2	0,1	10	95
30-34	1	0,05	5	100
<b>Total</b>	<b>20</b>	<b>1</b>	<b>100</b>	

source : D. ANDERSON, D. SWEENEY et T. WILLIAMS (2001)

### 3. Données quantitatives : regroupements quantitatifs



# 3. Données quantitatives : regroupements quantitatifs

Histogramme et notion de densité. Les histogrammes doivent représenter des densités, en particulier lorsque les classes ne sont pas d'amplitudes égales.

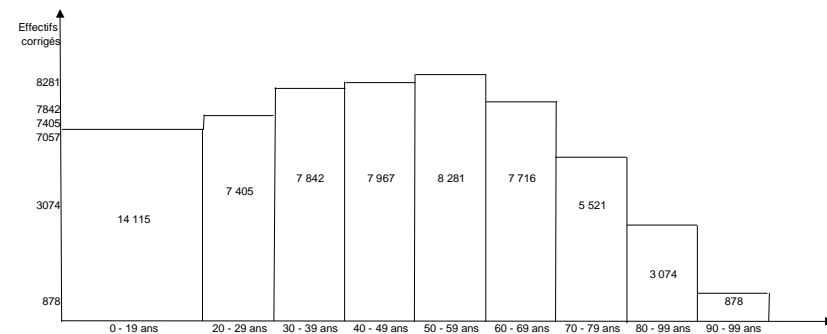
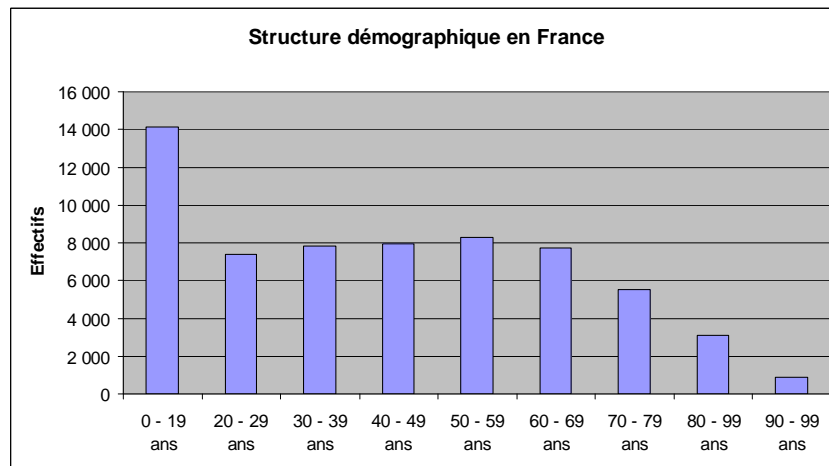
*Remarque : pas d'importance lorsque les classes sont d'amplitudes égales*

Structure démographique en France

âge ( $x_i$ )	nombre (en milliers) ( $n_i$ )	amplitude ( $a_i$ )	densité ( $d_i=n_i/a_i$ )	effectifs corrigés $nc_i = d_i \cdot \min(a_i)$
0 - 19 ans	14 115	20	705,75	7057,5
20 - 29 ans	7 405	10	740,5	7405
30 - 39 ans	7 842	10	784,2	7842
40 - 49 ans	7 967	10	796,7	7967
50 - 59 ans	8 281	10	828,1	8281
60 - 69 ans	7 716	10	771,6	7716
70 - 79 ans	5 521	10	552,1	5521
80 - 99 ans	3 074	10	307,4	3074
90 - 99 ans	878	10	87,8	878

source : E. BRESSOUD et J.C. KAHANE (2008) d'après INSEE, Projection à 2020, juillet 2006

### 3. Données quantitatives : regroupements quantitatifs



### 3. Données quantitatives :

## regroupements quantitatifs

#### Regroupement par superficie

Superficie	Amplitude de la classe	Effectif	Effectifs en %
[0 - 35200]	35 200	6	20
[41290 - 64569]	23279	6	20
[70263 - 110910]	40647	6	20
[131940 - 337030]	205090	6	20
[357021 - 9826830]	9469809	6	20
<b>Total</b>		<b>30</b>	<b>100</b>

#### Regroupement par superficie

Superficie	Amplitude de la classe	Effectif	Effectifs en %
[0 - 50 000]	50 000	11	36,67
[50 000 - 100 000]	50 000	6	20
[100 000 - 500 000]	400 000	10	33,33
[500 000 - 10 000 000]	9 500 000	3	10
<b>Total</b>		<b>30</b>	<b>100</b>

# 3. Données quantitatives : regroupements qualitatif

Regroupements par zone géographique

		nombre de pays	Fréquence	Emissions de gaz à effet de serre en 2003 (en millions de teq CO2)	Emissions de gaz à effet de serre en 2003 (en %)	PIB en 2003 (Milliards d'euros)	PIB en 2003 (en %)	Superficie (km2)	Superficie (en %)	Population (en millions)	Population (en %)
Europe	Zone Euro	15	55,56	3435,87	66,33	7515,02	74,34	2846743	64,96	321,30	64,63
	Hors Zone Eu	12	44,44	1743,96	33,67	2593,39	25,66	1535788	35,04	175,80	35,37
<b>Total</b>		<b>27</b>	<b>100</b>	<b>5179,83</b>	<b>100,00</b>	<b>10108,41</b>	<b>100,00</b>	<b>4382531,00</b>	<b>100,00</b>	<b>497,10</b>	<b>100,00</b>

---

# Chapitre 2 : Méthodes numériques permettant de résumer une série

---

---

# Plan

1. Statistiques résumant la tendance centrale
    - ❑ Moyennes
    - ❑ Médiane
    - ❑ Quantiles
    - ❑ mode
  2. Statistiques résumant la dispersion
    1. Variance
    2. écart-type
    3. coefficient de variation
-

---

# Introduction

Deux étudiants peuvent avoir des moyennes identiques mais avec des dispersion différentes

Un étudiant qui obtient une moyenne de 16/20, est-il un bon élève ?

Pour répondre à cette question, il faut connaître la moyenne médiane ou la répartition des notes.

# Statistiques résumant la tendance centrale : moyenne

Moyenne arithmétique simple :  $\bar{x} = \sum x_i / N$

Moyenne arithmétique pondérée :  $\bar{x} = \sum n_i x_i / N$  ou  $\bar{x} = \sum f_i x_i$

**Moyenne pondérée des salaires mensuelles**

Salaires ( $x_i$ )	$n_i$	$n_i x_i$	$f_i$	$f_i x_i$
1200	10	12000	0,13	160
1600	20	32000	0,27	426,67
2000	25	50000	0,33	666,67
2400	10	24000	0,13	320
2800	10	28000	0,13	373,33
<b>Total</b>	<b>75</b>	<b>146000</b>		<b>1946,67</b>
<b>Moyenne</b>		<b>1946,67</b>		<b>1946,67</b>

Source : B. PY (2007)

# Statistiques résumant la tendance centrale : moyenne

Moyenne avec des données groupées. On suppose que les données sont réparties de manière homogène à l'intérieur des classes.

Moyennes avec des données groupées

Durée des audits (jours) ( $x_i$ )	Nombre ( $n_i$ )	centre de classe ( $c_i$ )	$n_i c_i$
10-14	4	12	48
15-19	8	17	136
20-24	5	22	110
25-29	2	27	54
30-34	1	32	32
<b>Total</b>	<b>20</b>		<b>380</b>
<b>moyenne</b>			<b>19</b>

source : D. ANDERSON, D. SWEENEY et T. WILLIAMS (2001)

# Statistiques résumant la tendance centrale : moyenne

**Difficultés** : il est préférable de réaliser des moyennes sur des données brutes (quand cela est possible)

Regroupements par superficie

Superficie	Amplitude de la classe	Effectif	centre de classe	$n_i c_i$
[0 - 35200]	35 200	6	17 600	105600
[41290 - 64569]	23 279	6	52 929	317574
[70263 - 110910]	40 647	6	90 586	543516
[131940 - 337030]	205 090	6	234 485	1406910
[357021 - 9826830]	9 469 809	6	5 091 925	30551550
<b>Total</b>		<b>30</b>		<b>32925150</b>
<b>Moyenne</b>				<b>1 097 505</b>

Regroupements par superficie

Superficie	Amplitude de la classe	Effectif	centre de classe	Effectifs en %
[0 - 50 000]	50 000	11	25 000	275000
[50 000 - 100 000]	50 000	6	75 000	450000
[100 000 - 500 000]	400 000	10	300 000	3000000
[500 000 - 10 000 000]	9 500 000	3	5 250 000	15750000
<b>Total</b>		<b>30</b>		<b>19475000</b>
<b>Moyenne</b>				<b>649 166,70</b>

# Statistiques résumant la tendance centrale : moyenne

Superficie pour 30 pays

	Superficie ( <i>km</i> <sup>2</sup> )
Allemagne (1)	357021
Autriche	83858
Belgique	30528
Bulgarie	110910
Chypre	9250
Danemark	43094
Espagne	504762
Estonie	45225
Finlande	337030
France	643427
Grèce	131940
Hongrie	93030
Irlande	70263
Italie	301320
Lettonie	64569
Lituanie	35200
Luxembourg	2585
Malte	315
Pays-Bas	41526
Pologne	82931
Portugal	312665
République tchèque	78809
Roumanie	238391
Royaume-Uni	244820
Slovaquie	48845
Slovénie	20253
Suède	449964
Suisse	41290
Etats-Unis	9826830
Japon	377835
<b>Total de l'échantillon</b>	<b>14 628 486,0</b>
<b>Moyenne</b>	<b>487 616,2</b>

(1) : incluant l'ex-RDA à partir de 1991.

Source : EUROSTAT et INSEE

---

# Statistiques résumant la tendance centrale : moyenne

- Pour être significative, une moyenne doit être calculé sur un grand échantillon
- Elle est sensible aux valeurs extrêmes
- Ne suffit pas pour caractériser finement une série
- Il faut savoir quelles sont les variables dont on calcule la moyenne
  - Exemple : taux moyen d'absentéisme aux examens = 50%  
A quoi correspond un absent : absent à tous les examens ou absent a au moins un examen d'une même session.

---

# Statistiques résumant la tendance centrale : médiane

Médiane : correspond à la valeur centrale de la population

- Partage la population en 2.

50% de l'effectif se situe en dessous de la médiane et 50% de l'effectif se situe au dessus

Calcul : lorsque les données ont les mêmes effectifs pour chaque modalité (pays)

- Classer les données par ordre croissant
- Si l'effectif est impair, alors la médiane est la valeur centrale
- Si l'effectif est pair, alors la médiane est obtenue en faisant la moyenne des deux valeurs centrales.

# Statistiques résumant la tendance centrale : médiane

PIB pour UE

classement	Pays	PIB en 2003 (Milliards d'euros)
1	Malte	4,4214
2	Estonie	8,6926
3	Lettonie	9,9778
4	Chypre	11,785
5	Lituanie	16,4971
6	Bulgarie	17,7668
7	Slovénie	25,7359
8	Luxembourg	25,8343
9	Slovaquie	29,4856
10	Roumanie	52,613
11	Hongrie	74,5796
12	République tch	80,9241
13	Portugal	138,5821
<b>14</b>	<b>Irlande</b>	<b>139,4419</b>
15	Finlande	145,938
16	Grèce	171,4098
17	Danemark	188,5003
18	Pologne	191,6438
19	Autriche	223,3023
20	Belgique	274,726
21	Suède	275,657
22	Pays-Bas	476,945
23	Espagne	782,929
24	Italie	1335,3537
25	France	1594,814
26	Royaume-Uni	1647,0556
27	Allemagne (1)	2163,8
	<b>Union européenne</b>	<b>10 108,4</b>

(1) : incluant l'ex-RDA à partir de 1991.

Source : EUROSTAT et INSEE

PIB pour 30 pays

classement	Pays	PIB en 2003 (Milliards d'euros)
1	Malte	4,4214
2	Estonie	8,6926
3	Lettonie	9,9778
4	Chypre	11,785
5	Lituanie	16,4971
6	Bulgarie	17,7668
7	Slovénie	25,7359
8	Luxembourg	25,8343
9	Slovaquie	29,4856
10	Roumanie	52,613
11	Hongrie	74,5796
12	République tch	80,9241
13	Portugal	138,5821
14	Irlande	139,4419
<b>15</b>	<b>Finlande</b>	<b>145,938</b>
<b>16</b>	<b>Grèce</b>	<b>171,4098</b>
17	Danemark	188,5003
18	Pologne	191,6438
19	Autriche	223,3023
20	Belgique	274,726
21	Suède	275,657
22	Suisse	287,7538
23	Pays-Bas	476,945
24	Espagne	782,929
25	Italie	1335,3537
26	France	1594,814
27	Royaume-Uni	1647,0556
28	Allemagne (1)	2163,8
29	Japon	3743,5596
30	Etats-Unis	9689,5332
	<b>Total de l'écha</b>	<b>23 829,3</b>

Me = 158,22

(1) : incluant l'ex-RDA à partir de 1991.

Source : EUROSTAT et INSEE

---

# Statistiques résumant la tendance centrale : médiane

Calcul lorsque les effectifs ne sont pas les mêmes pour chaque observation

- Classer les observations par ordre croissant
- Calculer les fréquences cumulées
- Déterminer la médiane par interpolation linéaire

# Statistiques résumant la tendance centrale : médiane

Distribution des notes pour le restaurant Y

Note	Effectif	fréquence relative (%)	fréquence cumulée (%)	$f_i x_i$
1	2	4	4	0,04
2	6	12	16	0,24
<b>3</b>	<b>10</b>	<b>20</b>	<b>36</b>	0,6
<b>4</b>	<b>13</b>	<b>26</b>	<b>62</b>	1,04
5	19	38	100	1,9
<b>Total</b>	<b>50</b>	<b>100</b>		
<b>Moyenne</b>				<b>3,82</b>

source : D. ANDERSON, D. SWEENEY et T. WILLIAMS (2001)

$$\frac{Me - 3}{0,5 - 0,36} = \frac{4 - 3}{0,62 - 0,36} = \frac{1}{0,26} = 3,85$$

$$Me = 3,85 * 0,14 + 3 = 3,54$$

# Statistiques résumant la tendance centrale : médiane

Médiane avec des données par classe

Dépenses mensuelles en emplois à domicile

Dépense en euros	Effectifs	Fréquence en %	Fréquence cumulées (%)	centre de classe (c <sub>i</sub> )	f <sub>i</sub> c <sub>i</sub>
[300; 400[	5	2,38	2,38	350	8,33
[400; 500[	60	28,57	30,95	450	128,57
[500; 600[	15	7,14	38,09	550	39,29
<b>[600; 700[</b>	<b>95</b>	<b>45,24</b>	<b>83,33</b>	650	294,05
[700; 800[	30	14,29	97,62	750	107,14
[800; 1000[	5	2,38	100	900	21,43
<b>Total</b>	<b>210</b>	<b>100,00</b>			
<b>Moyenne</b>					<b>598,81</b>

Source : B. PY (2007)

$$\frac{Me - 600}{0,5 - 0,3809} = \frac{700 - 600}{0,8333 - 0,3809} = \frac{100}{0,4524} = 221,04$$

$$Me = 221,04(0,5 - 0,3809) + 600 = 626,326$$

---

# Statistiques résumant la tendance centrale : quantiles

Généralisent la médiane

- Quartiles : partagent les observations en 4 groupes égaux, chacun représentant 25% des observations
- Déciles : partagent les observations en 10 groupes égaux, chacun représentant 10% des observations
- Centiles : partagent les observations en 100 groupes égaux, chacun représentant 1% des observations

---

# Statistiques résumant la tendance centrale : quantiles

## Calcul

- Classer les données par ordre croissant
- Calculer l'indice

$$i = \frac{q}{100} N$$

Où  $q$  = quantile considéré  
 $N$  = nombre d'observations

- Si  $i$  n'est pas un nombre entier, on l'arrondit à l'entier supérieur
- Si  $i$  est un nombre entier, on détermine le quantile par la moyenne entre ce nombre et son supérieur ou par interpolation linéaire

---

# Statistiques résumant la tendance centrale : quantiles

## Exemple 1

avec le PIB des 30 pays : on cherche le 8<sup>ème</sup> décile, donc 80% des pays ont un PIB inférieur à ??

$$i = \frac{80}{100} 30 = 24$$

Le 8<sup>ème</sup> décile se trouve entre la 24<sup>ème</sup> et la 25<sup>ème</sup> position, soit entre l'Espagne et l'Italie

$$\text{Soit un PIB} = \frac{782,929 + 1335,3537}{2} = 1059,14$$

---

# Statistiques résumant la tendance centrale : quantiles

## **Exemple 2**

avec le PIB des 27 pays : on cherche le 1<sup>er</sup> quartile, donc 25% des pays ont un PIB inférieur à ??

$$i = \frac{25}{100} 27 = 6,75$$

Le 1<sup>er</sup> quartile correspond à la 7<sup>ème</sup> observation soit le PIB de la Slovénie

---

# Statistiques résumant la tendance centrale : mode

Le mode est la variable qui a l'effectif (ou la fréquence) le plus grand.

- Si la variable est qualitative ou quantitative discrète, le mode correspond à l'effectif (ou fréquence) maximal
- Si la variable est quantitative continue, on parle de classe modale et il faut calculer la valeur modale

*Remarque : Il peut ne pas exister de mode pour certaines séries  
(Données macroéconomiques des pays)*

Exemple 1 : pour les notes du restaurant Y, la note modale est 5

# Statistiques résumant la tendance centrale : mode

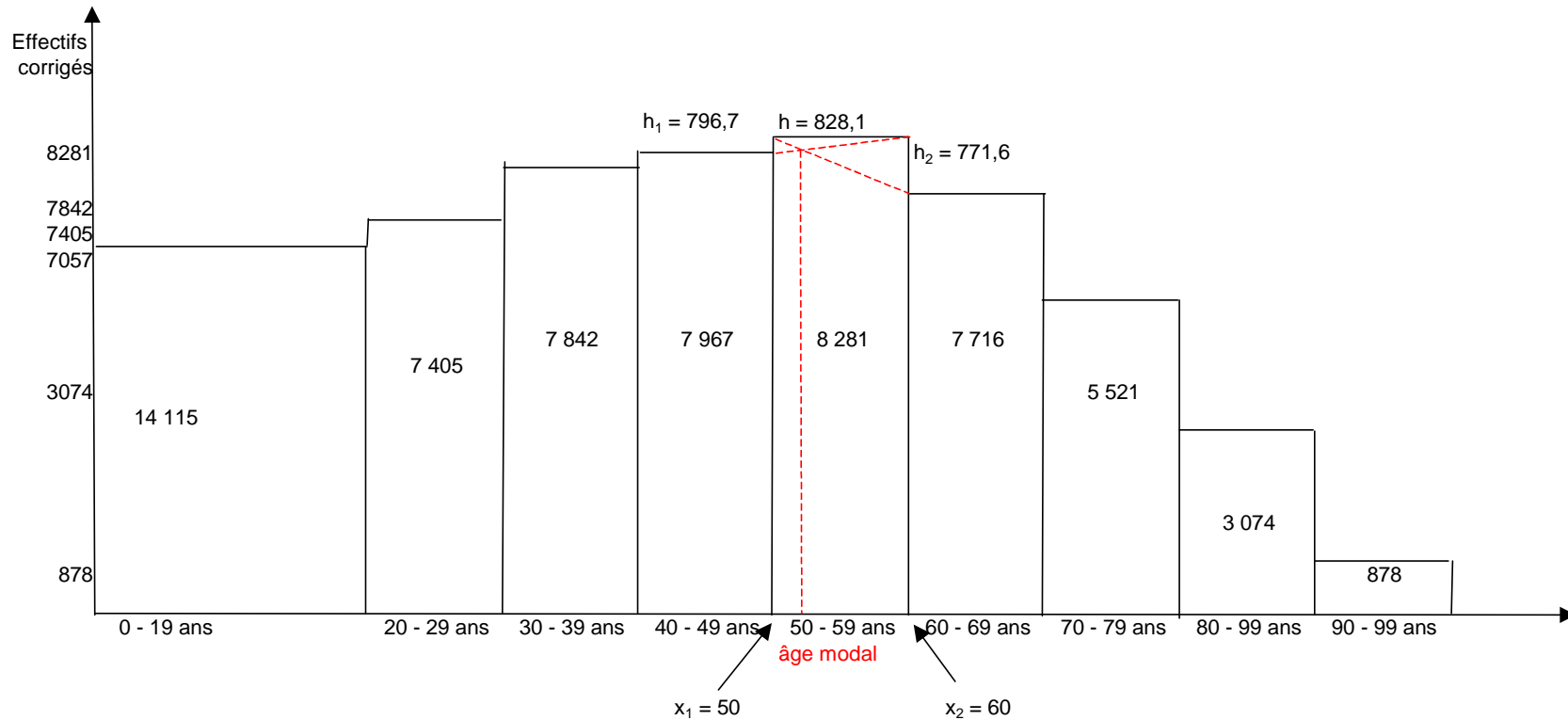
Exemple 2 : variables quantitatives continues

**Structure démographique en France**

âge ( $x_i$ )	nombre (en milliers) ( $n_i$ )	amplitude ( $a_i$ )	densité ( $d_i=n_i/a_i$ )	effectifs corrigés $nc_i = d_i * \min(a_i)$
0 - 19 ans	14 115	20	705,75	7057,5
20 - 29 ans	7 405	10	740,5	7405
30 - 39 ans	7 842	10	784,2	7842
40 - 49 ans	7 967	10	796,7	7967
50 - 59 ans	8 281	10	828,1	8281
60 - 69 ans	7 716	10	771,6	7716
70 - 79 ans	5 521	10	552,1	5521
80 - 99 ans	3 074	10	307,4	3074
90 - 99 ans	878	10	87,8	878

source : E. BRESSOUD et J.C. KAHANE (2008) d'après INSEE, Projection à 2020, juillet 2006

# Statistiques résumant la tendance centrale : mode

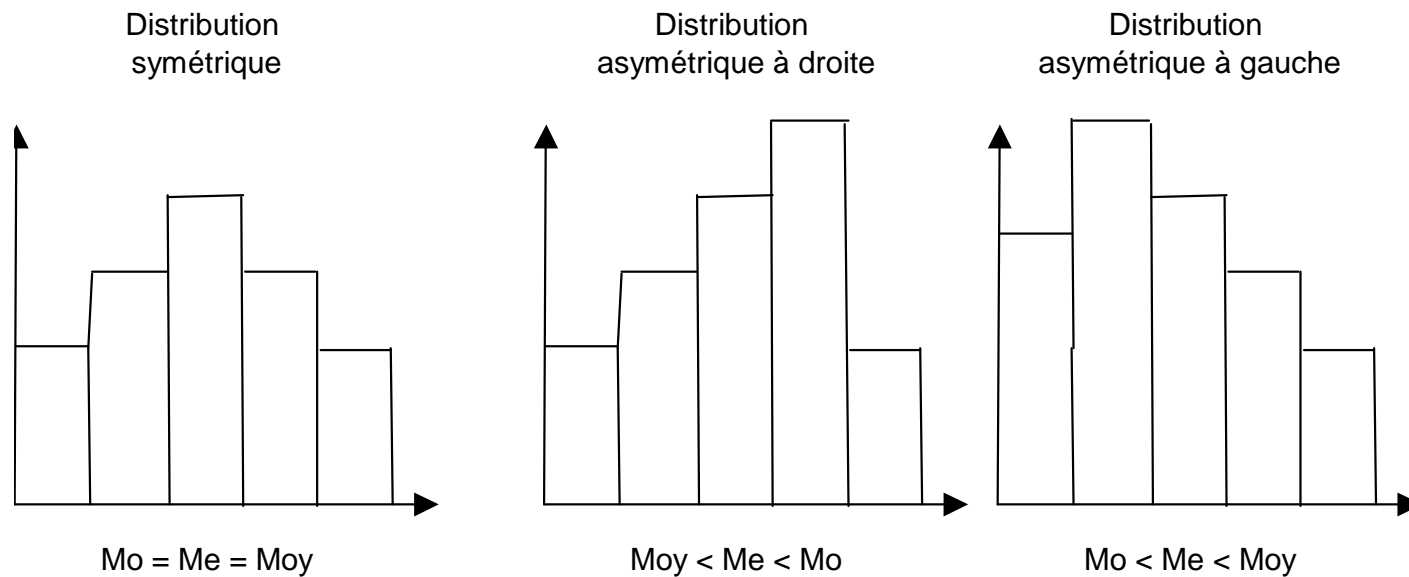


$$Mo = \frac{(h - h_1) x_2 + (h - h_2) x_1}{(h - h_1) + (h - h_2)}$$

$$Mo = \frac{(828,1 - 796,7)60 + (828,1 - 771,6)50}{(828,1 - 796,7) + (828,1 - 771,6)} = 53,57$$

# Statistiques résumant la tendance centrale : discussion

Moyenne, mode et médiane et forme d'une distribution



# Statistiques résumant la tendance centrale : discussion

Moyenne, mode et médiane : que choisir pour déterminer le centre d'une série ?

- Cela dépend du phénomène étudié et du message que l'on désire faire passer
- Il faut présenter la statistique la plus pertinente

Exemple 1 : moyenne ou position des étudiants

Exemple 2 : les salariés de l'entreprise A sont-ils mieux payés que ceux de l'entreprise B

Distribution de salaire dans 2 entreprises

	Entreprise A		Entreprise B	
	Salaires	Effectifs	Salaires	Effectifs
<b>Ouvriers</b>	1000	10	1500	15
<b>Cadres 1</b>	3000	2	2000	1
<b>Cadres 2</b>	5000	1	2500	1
<b>Total</b>	<b>9000</b>	<b>13</b>	<b>6000</b>	<b>17</b>
<b>Moyenne</b>	<b>1615</b>		<b>1588</b>	
<b>Mode</b>	<b>1000</b>		<b>1500</b>	

---

# Statistiques résumant la dispersion

La moyenne et/ou la médiane ne permettent pas d'apprécier la répartition des données.

- Valeur maximale et valeur minimale
- Intervalle de variation : valeur max. – valeur min.  
Pb : valeurs extrêmes peuvent être très différentes des autres valeurs
- Intervalle interquartile ou interdécile :  $Q3 - Q1$  ou  $D9 - D1$   
Délimitent la plage au sein de laquelle 50% ou 80% des valeurs sont regroupées  
Plus ces plages sont larges, plus les valeurs sont dispersées.  
Pb : ne pas prend en compte toutes les valeurs

---

# Statistiques résumant la dispersion

- Variance : somme des écarts à la moyenne, au carré

$$V(x) = \frac{1}{N} \sum_i n_i (x_i - \bar{x})^2 = \frac{1}{N} \sum_i n_i x_i^2 - \bar{x}^2$$

- Ecart-type : racine de la variance

$$\sigma_x = \sqrt{V(x)}$$

- Coefficient de variation : rapport entre l'écart-type et la moyenne

$$c_v = \frac{\sigma_x}{\bar{x}}$$

# Statistiques résumant la dispersion

**Notes des étudiants**

	Etudiant X	Etudiant Y	Etudiant Z
	0	7	12
	0	6	12
	0	15	12
	0	13	12
	20	4	12
	20	18	12
	20	20	12
	20	16	12
	20	12	12
	20	9	12
<b>Max</b>	<b>20</b>	<b>20</b>	<b>12</b>
<b>Min</b>	<b>0</b>	<b>4</b>	<b>12</b>
<b>intevalle de variation</b>	<b>20</b>	<b>16</b>	<b>0</b>
<b>moyenne</b>	<b>12</b>	<b>12</b>	<b>12</b>
<b>variance</b>	<b>96</b>	<b>26</b>	<b>0</b>
<b>écart-type</b>	<b>9,80</b>	<b>5,10</b>	<b>0</b>

# Statistiques résumant la dispersion : calculs

Distribution des notes pour le restaurant Y

Note	Effectif	$n_i x_i$	$n_i (x_i - \bar{X})^2$
1	2	2	15,90
2	6	12	19,87
<b>3</b>	<b>10</b>	30	6,72
<b>4</b>	<b>13</b>	52	0,42
5	19	95	26,46
<b>Total</b>	<b>50</b>	191	69,38
<b>Moyenne (X)</b>		<b>3,82</b>	
<b>variance</b>			<b>1,39</b>
<b>écart-type</b>			<b>1,18</b>
<b>coeff. Var.</b>			<b>0,31</b>

source : D. ANDERSON, D. SWEENEY et T. WILLIAMS (2001)

PIB pour 30 pays

Pays	PIB en 2003 (Milliards d'euros)	$(x_i - \bar{X})$	$(x_i - \bar{X})^2$
Malte	4,4214	-789,89	623921,80
Estonie	8,6926	-785,62	617192,52
Lettonie	9,9778	-784,33	615174,82
Chypre	11,785	-782,52	612343,20
Lituanie	16,4971	-777,81	604990,75
Bulgarie	17,7668	-776,54	603017,18
Slovénie	25,7359	-768,57	590704,01
Luxembourg	25,8343	-768,47	590552,77
Slovaquie	29,4856	-764,82	584954,24
Roumanie	52,613	-741,70	550112,38
Hongrie	74,5796	-719,73	518009,85
République tch	80,9241	-713,38	508917,46
Portugal	138,5821	-655,73	429977,26
Irlande	139,4419	-654,87	428850,41
Finlande	145,938	-648,37	420384,45
Grèce	171,4098	-622,90	388002,93
Danemark	188,5003	-605,81	367003,71
Pologne	191,6438	-602,66	363204,87
Autriche	223,3023	-571,01	326048,21
Belgique	274,726	-519,58	269966,09
Suède	275,657	-518,65	268999,49
Suisse	287,7538	-506,55	256597,78
Pays-Bas	476,945	-317,36	100719,66
Espagne	782,929	-11,38	129,50
Italie	1335,3537	541,05	292729,79
France	1594,814	800,51	640808,88
Royaume-Uni	1647,0556	852,75	727177,43
Allemagne (1)	2163,8	1369,49	1875506,67
Japon	3743,5596	2949,25	8698081,40
Etats-Unis	9689,5332	8895,22	79125020,51
<b>Total de l'écha</b>	<b>23 829,3</b>	<b>0,00</b>	<b>101999099,98</b>

moyenne (X)    **794,31**  
variance        **3399970,00**  
écart-type     **1843,90**  
coeff. Var.     **2,32**

# Statistiques résumant la dispersion : calculs avec des variables par classe

## Dépenses mensuelles en emplois à domicile

Dépense en euros	Effectifs	centre de classe ( $c_i$ )	$n_i c_i$	$n_i (c_i - X)^2$
[300; 400[	5	350	1750,00	309530,90
[400; 500[	60	450	27000,00	1328656,46
[500; 600[	15	550	8250,00	35735,54
<b>[600; 700[</b>	<b>95</b>	650	61750,00	248944,16
[700; 800[	30	750	22500,00	685756,80
[800; 1000[	5	900	4500,00	453578,51
<b>Total</b>	<b>210</b>		<b>125750,00</b>	<b>3062202,38</b>
<b>Moyenne (X)</b>			<b>598,81</b>	
<b>variance</b>				<b>14581,92</b>
<b>écart-type</b>				<b>120,76</b>
<b>coeff. Var.</b>				<b>0,58</b>

Source : B. PY (2007)

---

# Statistiques résumant la dispersion

Variance exprimée dans l'unité des données mais élevée au carré

⇒ Pour revenir à l'unité des données, on calcule l'écart-type

Mais ne permet pas de comparer les dispersions de 2 séries dont les unités sont différentes ⇒ coefficient de variation (nombre sans dimension)

# Conclusion

Données macroéconomiques pour les pays de l'UE à 27

	Emissions de gaz à effet de serre en 2003 (en millions de teq CO2)	PIB en 2003 (Milliards d'euros)	Population (en millions)	Densité moyenne (en hab./km2)	PIB/habitant (en milliers d'euros)	Pollution par habitant (en Teq CO2)	pollution/PIB (en kg eq CO2 par euro)
Allemagne (1)	1 030,1	2163,8	82,3	231	26,29	12,52	0,48
Autriche	93,3	223,3023	8,3	99	26,90	11,24	0,42
Belgique	146,3	274,726	10,6	347	25,92	13,80	0,53
Bulgarie	71,2	17,7668	7,7	69	2,31	9,25	4,01
Chypre	9,3	11,785	1,0	108	11,79	9,30	0,79
Danemark	73,8	188,5003	5,5	128	34,27	13,41	0,39
Espagne	410,1	782,929	45,3	90	17,28	9,05	0,52
Estonie	19,7	8,6926	1,3	29	6,69	15,15	2,27
Finlande	84,8	145,938	5,3	16	27,54	16,00	0,58
France	551,9	1594,814	63,6	99	25,08	8,68	0,35
Grèce	133,5	171,4098	11,2	85	15,30	11,92	0,78
Hongrie	80,6	74,5796	10,1	109	7,38	7,98	1,08
Irlande	68,6	139,4419	4,4	63	31,69	15,60	0,49
Italie	574,1	1335,3537	59,3	197	22,52	9,68	0,43
Lettonie	10,8	9,9778	2,3	36	4,34	4,72	1,09
Lituanie	21,0	16,4971	3,4	97	4,85	6,18	1,27
Luxembourg	11,7	25,8343	0,5	193	51,67	23,33	0,45
Malte	3,1	4,4214	0,4	1 270	11,05	7,65	0,69
Pays-Bas	216,3	476,945	16,4	395	29,08	13,19	0,45
Pologne	384,6	191,6438	38,1	459	5,03	10,09	2,01
Portugal	83,0	138,5821	10,7	34	12,95	7,76	0,60
République tchèque	145,5	80,9241	10,3	131	7,86	14,13	1,80
Roumanie	156,9	52,613	21,6	91	2,44	7,26	2,98
Royaume-Uni	658,9	1647,0556	61,0	249	27,00	10,80	0,40
Slovaquie	50,2	29,4856	5,4	111	5,46	9,30	1,70
Slovénie	19,8	25,7359	2,0	99	12,87	9,89	0,77
Suède	70,7	275,657	9,1	20	30,29	7,77	0,26
<b>Union européenne</b>	<b>5 179,8</b>	<b>10 108,4</b>	<b>497,1</b>	<b>113</b>	<b>20,33</b>	<b>10,42</b>	<b>0,51</b>

(1) : incluant l'ex-RDA à partir de 1991.

Source : EUROSTAT et INSEE

**Remarque : Attention aux calculs des totaux pour les 4 dernières colonnes (cela correspond aux moyennes de l'UE)**

# Conclusion

Données résumées pour les 27 pays de l'UE

	Emissions de gaz à effet de serre en 2003 (en millions de teq CO2)	PIB en 2003 (Milliards d'euros)	Population (en millions)	Densité moyenne (en hab./km2)	PIB/habitant (en milliers d'euros)	Pollution par habitant (en Teq CO2)	pollution/PIB (en kg eq CO2 par euro)
Moyenne	191,85	374,39	18,41	113,00	20,33	10,42	0,51
Valeur minimale	1030,10	2163,80	82,30	1269,84	51,67	23,33	4,01
Valeur maximale	3,06	4,42	0,40	15,73	2,31	4,72	0,26
Intervalle de variation	1027,04	2159,38	81,90	1254,12	49,36	18,61	3,75
Médiane	83,00	139,44	9,10	98,98	15,30	9,89	0,60
Q1	21,00	25,73	3,40	69,00	6,69	7,98	0,45
Q2	83,00	139,44	9,10	98,98	15,30	9,89	0,60
Q3	216,30	275,66	21,60	197,00	27,00	13,41	1,27
Intervalle interquartile	195,30	249,93	18,20	128,00	20,31	5,43	0,82
Ecart-type	246,25	582,41	22,81	240,63	12,14	3,78	0,89
Coefficient de variation	1,28	1,56	1,24	2,13	0,60	0,36	1,74

L'écart-type représente 213% de la moyenne pour la densité de population mais seulement 36% de la moyenne pour le PIB par habitant

Les données de densités de population sont 5,92 ( $2,13/0,36$ ) fois plus dispersées que celles des PIB par habitant

---

# Chapitre 3

## Indices et taux de croissance

---

# Plan

1. Comparaisons de données
2. Mesures de l'évolution des données
3. Les indices

# Comparaisons de données : Parts

Lorsqu'une variable est égale à la somme des ces composantes, on peut calculer la part de chaque composante par rapport à l'ensemble **pour une même date**

Chiffres d'affaires et nombre d'employés de l'hypermarché Machin pour différentes villes

Villes	CA en millions d'euros		Population (en milliers)
	2000	2008	2008
Brest	10000	11000	300
Caen	8000	9000	260
Nantes	20000	27000	800
Rennes	15 000	18000	500
<b>Total</b>	<b>53000</b>	<b>65000</b>	<b>1860</b>

Données fictives

---

# Comparaisons de données : Parts

$$\text{Part} = \text{CA}_{\text{ville}} / \text{Ca}_{\text{total}} * 100$$

Permet de visualiser l'évolution de la structure du chiffre d'affaire de cette entreprise

**Parts des Chiffres d'affaires de Machin  
(en %)**

<i>Villes</i>	<i>2000</i>	<i>2008</i>
Brest	18,87	16,92
Caen	15,09	13,85
Nantes	37,74	41,54
Rennes	28,30	27,69
<b>Total</b>	<b>100,00</b>	<b>100,00</b>

# Comparaisons de données : Ecart relatif et absolu

Permet de comparer des variables à une même date pour des individus différents

Ecart absolu = valeur  $i$  – valeur  $j$

Ecart relatif =  $((\text{valeur } i - \text{valeur } j) / \text{valeur } j) * 100$

=  $(\text{valeur } i / \text{valeur } j - 1) * 100$

**Comparaisons des CA**

Villes	écart absolu (en millions d'euros)	écart relatif (en %)
Rennes - Brest	5 000	50
Brest - Rennes	-5 000	-33,33

Remarque : Attention au sens du calcul de l'écart relatif

# Comparaisons de données : Ratio

Rapport significatif entre 2 variables. Permet d'affiner l'analyse à **une même date**

CA et CA/population

	<i>CA (en millions d'euros)</i>	<i>Rang</i>	<i>Population (en milliers)</i>	<i>CA/population (en millions d'euros)</i>	<i>Rang</i>
Brest	11000	3	300	36,67	1
Caen	9000	4	260	34,62	3
Nantes	27000	1	800	33,75	4
Rennes	18000	2	500	36,00	2
<b>Total</b>	<b>65000</b>		<b>1860</b>	<b>34,95</b>	

---

# Mesures de l'évolution

Mesure l'évolution d'une variable **entre deux dates différentes pour un même individu**

Notations :

$V_0$  : valeur à la date  $t = 0$

$V_1$  : valeur à la date  $t = 1$

$V_t$  : valeur à la date  $t$

$g_t$  : taux de croissance entre les dates  $t$  et  $t+1$

Variation absolue =  $V_t - V_0$

Variation relative = taux de croissance

$$= ((V_t - V_0) / V_0) * 100$$

$$= (V_t / V_0 - 1) * 100$$

# Mesures de l'évolution

<i>Villes</i>	<i>CA (en millions d'euros)</i>		<i>Evolutions</i>	
	<i>2000</i>	<i>2008</i>	<i>Ecart absolu(en millions d'euros)</i>	<i>écart relatif (en %)</i>
Brest	10000	11000	1000	10
Caen	8000	9000	1000	12,5
Nantes	20000	27000	7000	35
Rennes	15 000	18000	3000	20
<b>Total</b>	<b>53000</b>	<b>65000</b>	<b>12000</b>	<b>22,64</b>

---

# Mesures de l'évolution : taux de croissance

$$V_{2008} = (1+g)^*V_{2000}$$

$$V_{2000} = V_{2008} / (1+g)$$

Attention : Les taux de croissance ne sont pas additifs

Points de croissance = différence entre deux taux de croissance

Le taux de croissance de Caen est 2,5 **points** plus élevé que le taux de croissance de Brest

---

# Mesures de l'évolution : taux de croissance

Taux de croissance d'un produit

$$\Pi = x \cdot y$$

$$g_{\Pi} = (1+g_x)(1+g_y) - 1$$

Taux de croissance d'un quotient

$$Q = x/y$$

$$g_Q = (1+g_x)/(1+g_y) - 1$$

Approximation : Pour de faibles taux de croissance (< 10%)

$$g_{\Pi} \approx g_x + g_y$$

$$g_Q \approx g_x - g_y$$

# Mesures de l'évolution : taux de croissance annuel moyen

Produit intérieur brut aux prix de marché (en valeur)

	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Niveau	1315,26	1367,97	1441,37	1497,17	1548,56	1594,81	1660,19	1726,07	1807,46	1892,24
taux de croissance		1999/1998	2000/1999	2001/2000	2002/2001	2003/2002	2004/2003	2005/2004	2006/2005	2007/2006
		4,01	5,37	3,87	3,43	2,99	4,10	3,97	4,72	4,69

On cherche le taux de croissance identique pour chaque période qui donnerait la même évolution sur la période

$$V_1 = (1+g)*V_0$$

$$V_2 = (1+g)*V_1 = (1+g)^2 *V_0$$

$$V_3 = (1+g)*V_2 = (1+g)^3 *V_0$$

...

$$V_9 = (1+g)^9 *V_0 \Rightarrow g = (V_9/V_0)^{1/9} - 1$$

---

## Mesures de l'évolution : taux de croissance annuel moyen

$$g = (1892,24/1315,26)^{1/9} - 1 = 0,0412$$

Le taux de croissance annuel moyen est de 4,12%

# Mesures de l'évolution : contribution à la croissance

Question : quelle la contribution de chaque ville à la croissance du CA de l'hypermarché Machin ? Ou quel est le magasin qui entraîne le plus la croissance du groupe ?

$$CA_{\text{total}} = CA_{\text{Brest}} + CA_{\text{Caen}} + CA_{\text{Nantes}} + CA_{\text{Rennes}}$$

$$g_{CA_{\text{total}}} = \text{Part}_{CA_{\text{Brest}2000}} * g_{CA_{\text{Brest}}} + \text{Part}_{CA_{\text{Caen}2000}} * g_{CA_{\text{Caen}}} + \text{Part}_{CA_{\text{Nantes}2000}} * g_{CA_{\text{Nantes}}} + \text{Part}_{CA_{\text{Rennes}2000}} * g_{CA_{\text{Rennes}}}$$

**Contribution à la croissance du CA de Machin**

Villes	CA en millions d'euros		Parts	Taux de croissance	Contribution
	2000	2008			
Brest	10000	11000	18,87	10,00	1,89
Caen	8000	9000	15,09	12,50	1,89
Nantes	20000	27000	37,74	35,00	13,21
Rennes	15 000	18000	28,30	20,00	5,66
<b>Total</b>	<b>53000</b>	<b>65000</b>		<b>22,64</b>	<b>22,64</b>

---

# Les indices

De nombreuses variables sont exprimées sous forme d'indices

Un indice évalue une variation et non un niveau

## Exemple

L'indice du taux de change €/ \$ en 2008 base 100 en 2002 est 160,  
alors l' s'est apprécié de 60% par rapport au \$

---

# Les indices élémentaires

Un indice est un rapport de la même variable prise à deux dates différentes ou lieux distincts

## Définition

*Indice élémentaire de la variable  $G$ , à la date  $t$ , base 1 en  $t = 0$ , est  $I_{t/0} = G_t/G_0$*

*Indice élémentaire de la variable  $G$ , à la date  $t$ , base 100 en  $t = 0$ , est  $I_{t/0} = G_t/G_0 * 100$*

*Indice élémentaire chaîné de la variable  $G$ , à la date  $t$ , base 100 en  $t = t-1$ , est  $I_{t/t-1} = G_t/G_{t-1} * 100$*

# Les indices élémentaires

Produit intérieur brut aux prix de marché (en valeur)

	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Niveau	1315,26	1367,97	1441,37	1497,17	1548,56	1594,81	1660,19	1726,07	1807,46	1892,24
taux de croissance		4,01	5,37	3,87	3,43	2,99	4,10	3,97	4,72	4,69
Indice (base 100 en 1998)	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
	100	104,01	109,59	113,83	117,74	121,25	126,22	131,23	137,42	143,87
Indice (base 100 en 2002)	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
	84,93	88,34	93,08	96,68	100,00	102,99	107,21	111,46	116,72	122,19
Indice chaîné (base 100 en t-1)	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
	-	104,01	105,37	103,87	103,43	102,99	104,10	103,97	104,72	104,69

Base 100 en 1998 : entre 1998 et 2007, les PIB en valeur a augmenté de 43,87%

Base 100 en 2002 : entre 2002 et 2005, le PIB en valeur a augmenté de 11,46%

*Attention : on ne connaît la progression que par rapport à l'année de base*

Taux de croissance entre 2000 et 2001  $\neq 113,83 - 109,59 = 4,24\%$

Voir indices chaînés

---

# Les indices élémentaires : propriétés

## Circularité

Base 1:  $I_{t_2/t_0} = I_{t_2/t_1} * I_{t_1/t_0}$

Base 100:  $I_{t_2/t_0} = I_{t_2/t_1} * I_{t_1/t_0} * 100$

Exemple :  $I_{2001/2000} = I_{2001/1998} / I_{2000/1998} * 100$

$$I_{2001/2000} = 113,83/109,59 = 103,87$$

Donc les PIB en valeur a augmenté de 3,87% entre 2000 et 2001

## Réversibilité

$$I_{t_1/t_0} = 1 / I_{t_0/t_1}$$

# Les indices synthétiques

Comment synthétiser l'évolution simultanée de plusieurs variables.

**Prix et quantités consommées du café et du sucre**

	café			sucre			<b>dépense totale</b>
	Prix	Quantité	dépense	Prix	Quantité	Dépense	
2000	0,8	100	80	0,2	90	18	<b>98</b>
2008	1,4	120	168	0,5	70	35	<b>203</b>

Possibilité de calculer les indices élémentaires pour chaque variable (4 indices)

**Indices élémentaires du café et du sucre base 100 en 2000**

	café	sucre
2000	100	100
2008	210	194,44

⇒ Construction d'indices synthétiques

# Les indices synthétiques

Indice de valeur :

$$I_{t/0} = \frac{\sum_i p_t^i q_t^i}{\sum_i p_0^i q_0^i} 100$$

**Indices de valeur de la consommation  
de café et de sucre base 100 en 2000**

2000	100
2008	207,14

Indice mesure l'évolution des prix et des quantités

⇒ Calculs d'indices qui fixent les quantités et donc mesure uniquement l'évolution des prix

# Les indices synthétiques : Indice de Laspeyres

Indice de Laspeyres des prix fixe les quantités à l'année de départ (2000)

⇒ Seuls les prix évoluent

$$L_{t/0} = \frac{\sum_i p_t^i q_0^i}{\sum_i p_0^i q_0^i} 100$$

**Indice de Laspeyres base 100 en 2000**

Dépense 2000 prix 2000*quantité 2000	Dépense 2008 Prix 2008*quantité 2000	Indice de Laspeyre
98	185	188,78

Indice de Laspeyres = moyenne pondérée des indices élémentaires par les coefficients budgétaires calculés à la date de la base

# Les indices synthétiques : Indice de Paasche

Indice de Paasche des prix fixe les quantités à l'année finale ou année courante (2008)

$$P_{t/0} = \frac{\sum_i p_t^i q_t^i}{\sum_i p_0^i q_t^i} 100$$

**Indice de Paasche base 100 en 2000**

Dépense 2000 prix 2000*quantité 2008	Dépense 2008 Prix 2008*quantité 2008	Indice de Paasche
110	203	184,55

---

## Indices : remarques finales

Possibilités de calculer des indices de quantités en fixant cette fois les prix

L'INSEE utilise l'indice de Lapeyres pour calculer l'indice des prix à la consommation

---

# Chapitre 4

## Corrélation et liaisons entre des variables

---

# Introduction

Jusqu'à présent, nous avons utilisé des méthodes pour résumer les données pour une variable à un moment donné ou dans le temps.

Dans ce chapitre, nous étudierons le croisement de deux ou plusieurs variables (statistiques bi ou pluridimensionnelles).

Le but du croisement de variables est la recherche de l'existence d'un lien de dépendance entre ces variables ou d'une liaison

Exemples :

Existe-t-il un lien entre le PIB et les émissions de gaz à effet de serre ?

Existe-t-il un lien entre la vente de certains produits et l'âge ou le sexe des consommateurs ?

Existe-t-il un lien entre le salaire et l'âge des salariés ?

---

# Introduction

On cherche un lien de dépendance ou d'indépendance entre des variables statistiques

Si ce lien existe, comment le modéliser ?

Attention : la question de la liaison entre deux variables est différente de la question du sens de la causalité.

Exemple :

Est-ce le prix qui détermine la demande ou la demande qui explique le niveau des prix ?

---

# Plan

- Etude des liaisons statistiques pour des données quantitatives
  - Analyse graphique
  - La covariance et le coefficient de corrélation
  - La régression
  
- Etude des liaisons statistiques pour des données qualitatives
  - Présentation des tableaux croisés
  - Les tableaux de contingences
  - Fréquences conditionnelles
  - Indépendance des variables (test du Khi-deux)

# Données quantitatives : nuages de points

CA et spots publicitaires pour le magasin Truc

Semaines	Nombres de spots publicitaires	CA en centaines de dollars
1	2	50
2	5	57
3	1	41
4	6	54
5	5	54
6	1	38
7	6	63
8	3	48
9	4	59
10	7	65

Source : adapté de Anderson et alii ( 2001)

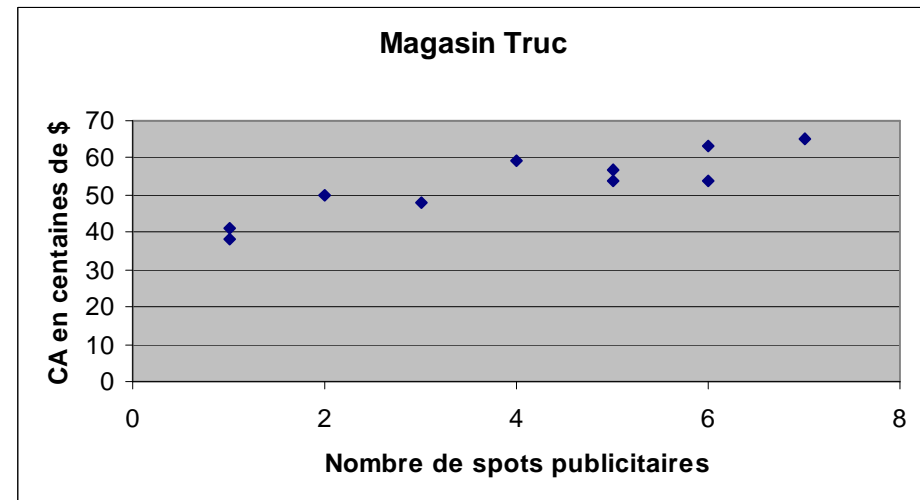
Question : existe-t-il une liaison statistique entre le nombre de spots et le CA ?

Le CA et le nombre de spots évoluent-ils de manière concomitante ?

# Données quantitatives : nuages de points

Un représentation graphique du nuage de points (ou diagramme de corrélation) permet :

- D'apprécier l'existence ou non d'une éventuelle liaison
- De déterminer la forme de la liaison



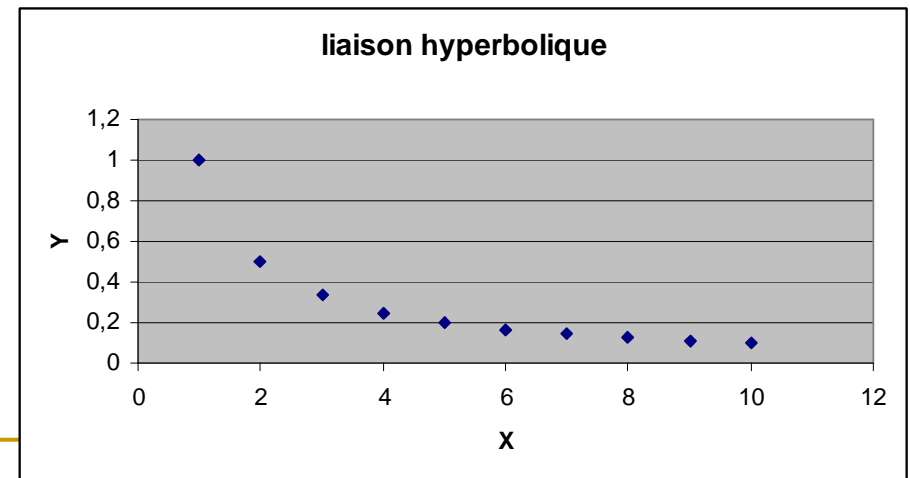
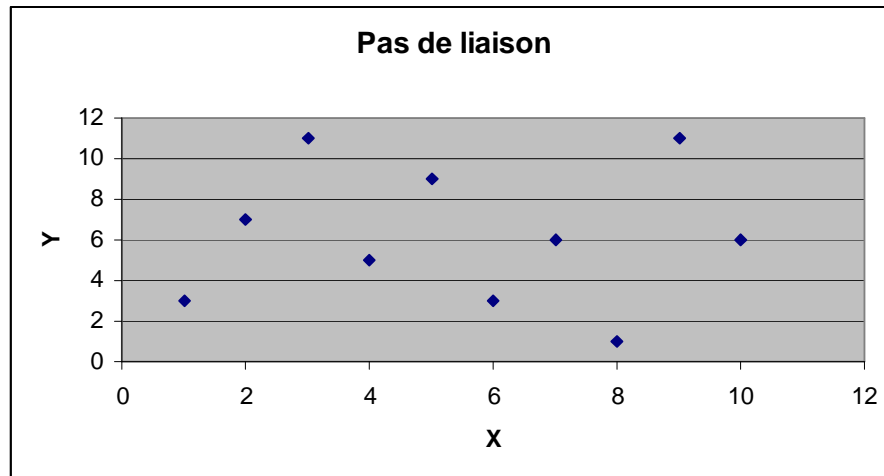
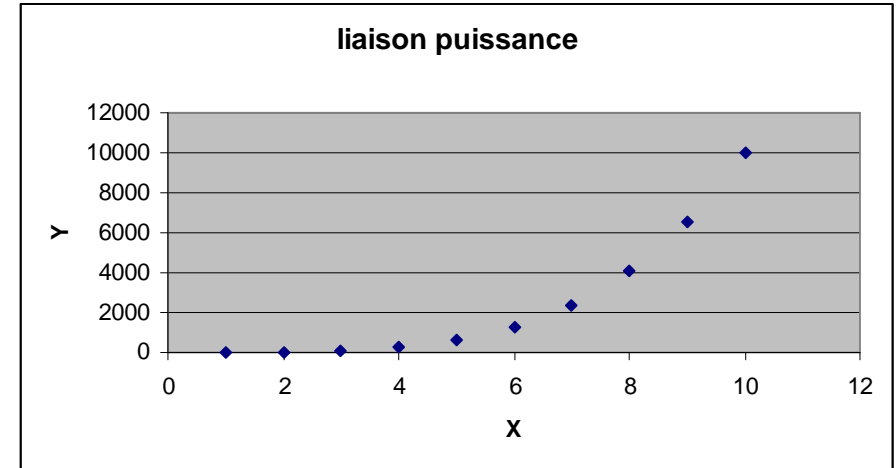
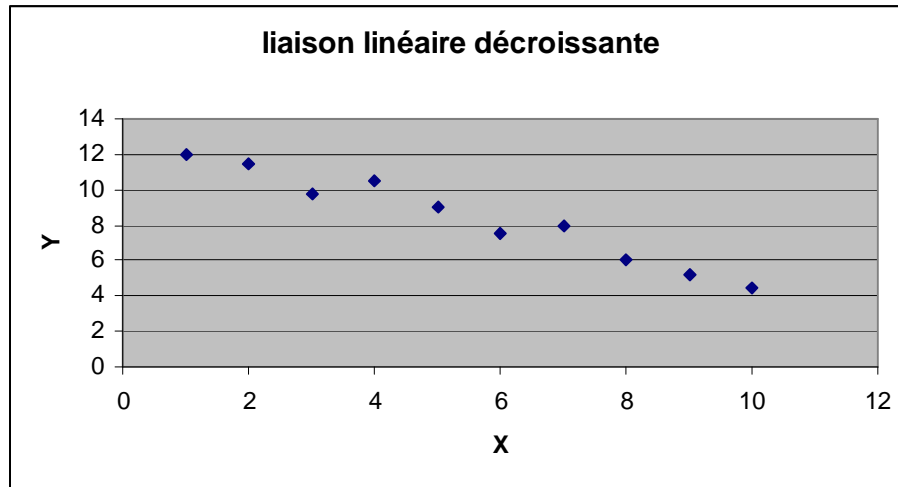
---

# Données quantitatives : nuages de points

La forme du nuage de point suggère les interprétations suivantes :

- Il existe une liaison entre les 2 variables : si le nombre de spots varie alors le CA a tendance à varier aussi
- Cette liaison est linéaire : les points sont à peu près alignés sur une droite
- Cette liaison est positive : plus le nombre de spots s'accroît, plus le CA augmente.

# Nuages de points : formes de liaison



---

# Covariance

Pour le magasin, le nuage de points montre que les variables ont tendance à covarier (varier ensemble)

⇒ Construction d'un indicateur qui mesure la variabilité conjointe des 2 variables.

- Mesure descriptive de la relation entre les 2 variables
- Mesure les fluctuations simultanées de chaque variable par rapport à sa moyenne

# Covariance : calculs

$$COV(X, Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{N}$$

$$COV(X, Y) = \frac{1}{N} \sum_i x_i y_i - \bar{x} \bar{y}$$

COV (X, Y) = moyenne du produit XY – produit des moyennes de X et de Y

Calcul de la covariance pour le magasin Truc

Semaines	Nombres de spots publicitaires (X)	CA en centaines de dollars (Y)	XY
1	2	50	100
2	5	57	285
3	1	41	41
4	6	54	324
5	5	54	270
6	1	38	38
7	6	63	378
8	3	48	144
9	4	59	236
10	7	65	455
<b>Moyenne</b>	<b>4</b>	<b>52,9</b>	<b>227,1</b>

$$\text{Covariance} = 227,1 - 4 * 52,9 = 15,5$$

---

# Covariance : interprétation

Covariance  $> 0 \Rightarrow$  les variables ont tendance à varier dans le même sens

Covariance  $< 0 \Rightarrow$  les variables ont tendance à varier en sens opposée

$\Rightarrow$  Plus la valeur ( $>0$  ou  $<0$ ) de la covariance est élevée plus la relation entre les variables est forte

$\Rightarrow$  S'il n'y a pas de tendance à la croissance ou à la décroissance entre les variables covariance nulle

☞ La covariance est un indicateur de relation linéaire entre les variables

$\Rightarrow$  Covariance = 0 peut signifier une relation non linéaire.

# Coefficient de corrélation linéaire

Covariance dépend des unités des variables  $\Rightarrow$  coefficient de corrélation linéaire.

Coefficient de corrélation linéaire

$$r = \frac{COV(X,Y)}{\sigma_x \sigma_y}$$

$$r = \frac{15,5}{2,049 * 8,37} = 0,903$$

- $-1 < r < 1$
- Si  $r = 1$  ou  $r = -1$  alors points parfaitement alignés

---

# Régression linéaire

Il s'agit de caractériser quantitativement le lien entre les deux variables.

Seule situation envisagée : le nuage de points suggère une liaison linéaire :

$$\Rightarrow y = ax + b$$

En connaissant l'équation de la droite qui résume la relation, il est possible de faire des prévisions

Remarque : attention à la véracité statistique de ces prévisions lorsqu'on sort de l'intervalle de l'échantillon

# Régression linéaire

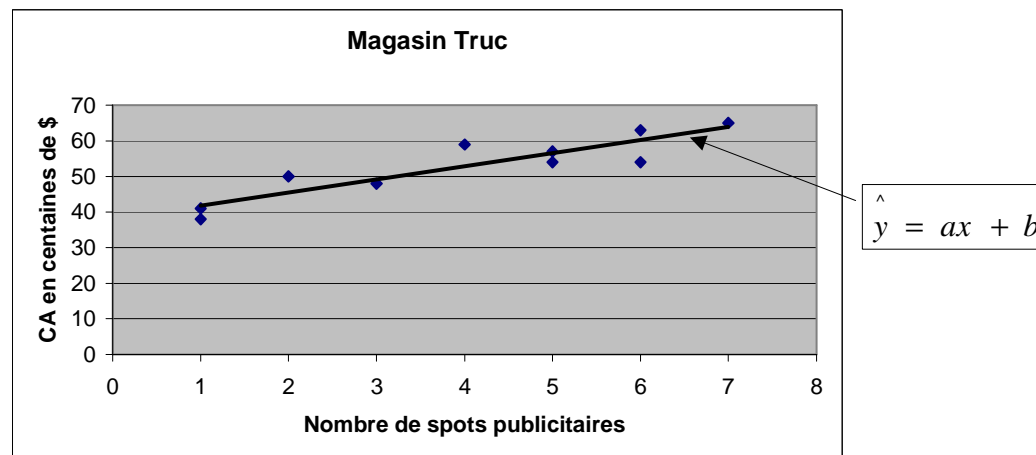
On cherche donc à estimer la droite qui s'ajuste le mieux au nuage de point

Notation

$y$  = vraies valeurs de la valeur de variable  $y$  c'est la variable expliquée

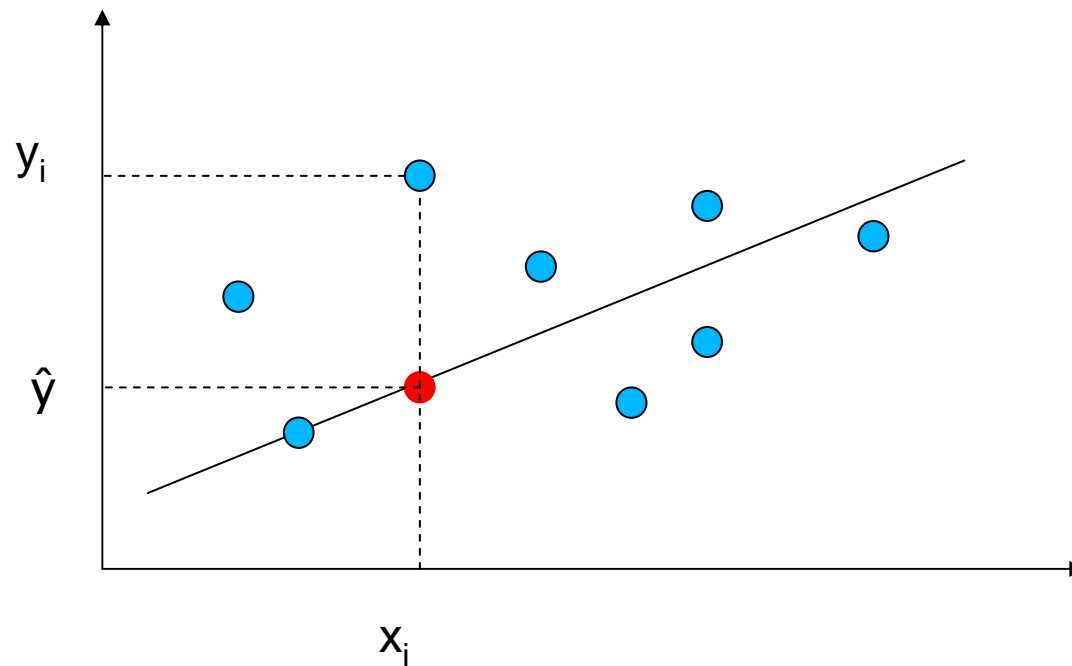
$\hat{y}$  = valeurs de la variables  $y$  obtenues à l'aide du modèle

$x$  = variable dépendante ou variable explicative



# Régression linéaire

Méthodologie : minimisation de la somme des carrés des écarts entre la véritable valeurs de  $y_i$  et son estimation



# Régression linéaire

La droite de régression  
a pour équation

$$a = \frac{COV(X,Y)}{Var(X)}$$

$$b = \bar{y} - a \bar{x}$$

Calcul de la droite de régression pour le magasin Truc

Semaines	Nombres de spots publicitaires (X)	CA en centaines de dollars (Y)	XY	X <sup>2</sup>
1	2	50	100	4
2	5	57	285	25
3	1	41	41	1
4	6	54	324	36
5	5	54	270	25
6	1	38	38	1
7	6	63	378	36
8	3	48	144	9
9	4	59	236	16
10	7	65	455	49
<b>Total</b>	<b>40</b>	<b>529</b>	<b>2271</b>	<b>202</b>
<i>Moyenne</i>	<b>4</b>	<b>52,9</b>	<b>227,1</b>	

$$Cov(X,Y) = 227,1 - 4*52,9 = 15,5$$

$$Var(X) = 202/10 - 4^2 = 4,2$$

$$a = 15,5/4,2 = 3,69$$

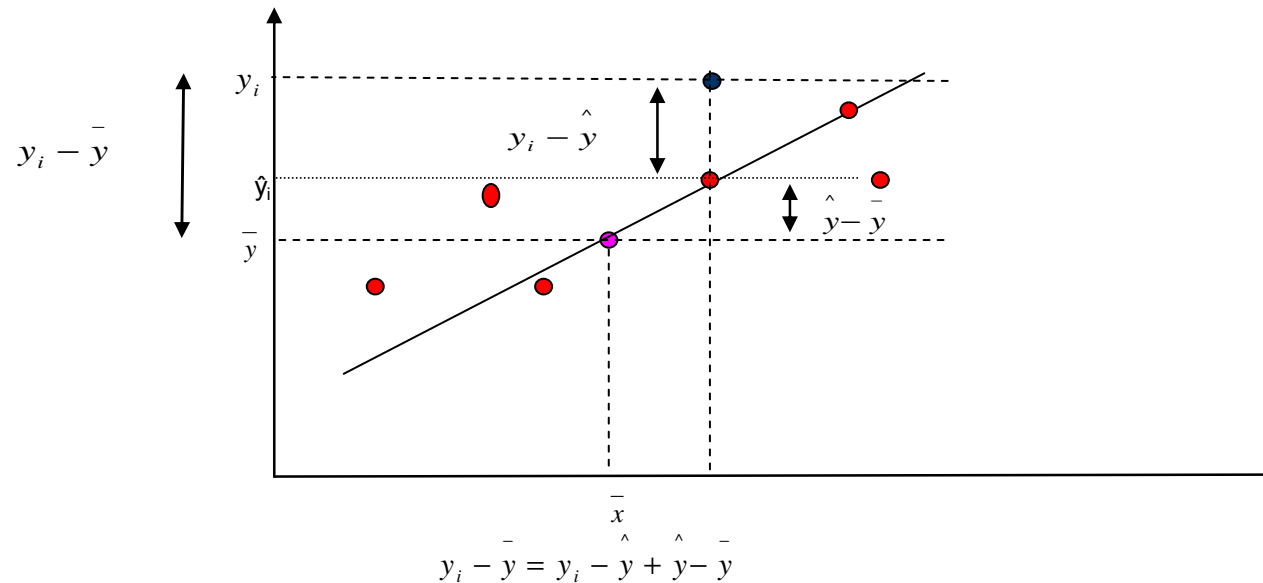
$$b = 52,9 - 3,69*4 = 38,14$$

$$\hat{y} = 3,69x + 38,14$$

# Régression linéaire : coefficient de détermination

Cette droite explique-t-elle de façon satisfaisante les variations de  $y$  (ou la variance de  $y$ )

La droite de régression passe par la covariance  $\Rightarrow$  moy  $(\hat{y}) = \bar{y}$



on montre que  $\left( y_i - \bar{y} \right)^2 = \left( \hat{y} - \bar{y} \right)^2 + \left( y_i - \hat{y} \right)^2 \Rightarrow SCT = SCE + SCR$

# Régression linéaire : coefficient de détermination

Calcul de la covariance pour le magasin Truc

Semaines	Nombres de spots publicitaires (X)	CA en centaines de dollars (Y)	$\hat{Y}$	$(Y - \hat{Y})$	$(Y - m_y)^2$	$(\hat{Y} - m_{\hat{y}})^2$	$(Y - \hat{Y})^2$
1	2	50	45,52	4,48	8,41	54,48	20,08
2	5	57	56,59	0,41	16,81	13,62	0,17
3	1	41	41,83	-0,83	141,61	122,58	0,69
4	6	54	60,28	-6,28	1,21	54,48	39,45
5	5	54	56,59	-2,59	1,21	13,62	6,71
6	1	38	41,83	-3,83	222,01	122,58	14,66
7	6	63	60,28	2,72	102,01	54,48	7,39
8	3	48	49,21	-1,21	24,01	13,62	1,46
9	4	59	52,90	6,10	37,21	0,00	37,21
10	7	65	63,97	1,03	146,41	122,58	1,06
<b>Total</b>	<b>40</b>	<b>529</b>			<b>700,90</b>	<b>572,02</b>	<b>128,88</b>
<b>Moyenne</b>	<b>4</b>	<b>52,9</b>			<b>SCT</b>	<b>SCE</b>	<b>SCR</b>

$$a = \frac{15,5}{4,2} = 3,69$$

$$b = 52,9 - 3,69 \cdot 4 = 38,14$$

$$\hat{y} = 3,69x + 38,14$$

$$SCT = 572,02 + 128,88 = 700,90$$

$$R^2 = \frac{SCE}{SCT}$$

$$R^2 = \frac{572,02}{700,9}$$

$$R^2 = 81,61$$

---

# Régression linéaire : coefficient de détermination

$R^2$  représente la part de la variabilité de  $Y$  « expliquée » par la droite de régression.

$$R^2 \leq 1$$

Si les observations sont parfaitement alignées, il n'y a pas de différence entre  $y$  et  $\hat{y} \Rightarrow$  pas de résidu  $\Rightarrow$   $SCT = SCE \Rightarrow R^2 = 1$

Donc  $R^2$  exprime la qualité du modèle. Plus est proche de 1, meilleure est la qualité du modèle linéaire

Ici le nombre de spots publicitaires « explique » 81,61% de la dispersion des CA

Remarque :  $R^2 = r^2$ , uniquement pour un modèle linéaire

---