

# Statistiques

1

---

---

---

---

---

---

---

---

## 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

2

---

---

---

---

---

---

---

---

## 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

3

---

---

---

---

---

---

---

---

## Introduction

Qu'est ce que la statistique ?

4

---

---

---

---

---

---

---

---

## 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)

5

---

---

---

---

---

---

---

---

## 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

6

---

---

---

---

---

---

---

---

## 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...

7

---

---

---

---

---

---

---

---

## 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

8

---

---

---

---

---

---

---

---

## 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

9

---

---

---

---

---

---

---

---

## 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

10

---

---

---

---

---

---

---

---

## 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

11

---

---

---

---

---

---

---

---

## 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)

12

---

---

---

---

---

---

---

---

## 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...)

13

## 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 ton CO2)	PIB en 2003 (Milliards d'euros)	Superficie (km <sup>2</sup> )	Population (en millions)	Population urbaine (en %)	Monnaie
Allemagne (1)	1 030,1	2 063,8	35 7021	82,1	75	euro
Autriche	93,3	252,3025	83856	8,3	67	euro
Belgique	146,3	279,726	30523	10,6	97	euro
Bulgarie	71,2	17,7668	110910	7,7	71	Lev
Cyprus	9,4	11,085	9250	1,0	62	euro
Danemark	73,8	188,5001	43064	5,5	72	Couronne danoise
Espagne	418,1	782,925	504762	45,3	77	euro
Estonie	19,7	8,6026	45223	1,3	69	Couronne estonienne
Finlande	84,8	145,978	337030	5,3	62	euro
France	551,9	1594,814	643427	63,6	77	euro
Grèce	123,5	171,4098	131940	11,7	59	euro
Hongrie	80,6	14,3796	93039	10,1	65	Florint
Irlande	68,6	139,4419	70281	4,4	60	euro
Italie	574,1	1 135,3511	301120	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	2583	0,5	83	euro
Malte	3,1	4,2314	315	0,4	85	euro
Pays-Bas	216,3	478,945	41526	16,4	65	euro
Pologne	384,6	191,6438	32911	38,1	62	Zlot
Portugal	83,0	138,5821	312666	10,7	55	euro
République tchèque	145,5	300,241	78809	10,3	74	Couronne tchèque
Roumanie	156,9	52,613	238391	21,6	55	Leu
Royaume-Uni	658,9	167,056	244820	61,0	69	Livre sterling
Slovaquie	50,2	29,4856	58847	5,4	56	Couronne slovaque
Slovenie	19,8	25,7389	20253	2,0	49	euro
Suède	70,7	274,663	449664	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
Israël	6 890,8	9689,4332	9604630	680,2	79	Sheqel
Israël	1 376,4	144,5596	67783	12,7	79	Sheqel
<b>Total de l'échantillon</b>	<b>13 465,4</b>	<b>31 829,4</b>	<b>14 618 496,0</b>	<b>934,8</b>	-	-

(1) Incluant l'ex-REDA à partir de 1991.  
Source: EUROSTAT et INSEE.

14

## Exemple 1

Population = 30 pays ou 30 unités statistiques  
Cette population est décrite par 6 critères

15

## Exemple 2 : tableau croisé

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

	Cursus Licence	Cursus Master	Cursus Doctorat	Effectif total
	<i>Effectif</i>	<i>Effectif</i>	<i>Effectif</i>	
Droit, sciences politiques	106690	64064	8371	179125
Sciences économiques, gestion (hors AES)	75544	56389	4535	136474
Administration économique et sociale (AES)	30962	7097	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)

16

## 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

17

## 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

18

## 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

19

---

---

---

---

---

---

---

---

## 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...)

20

---

---

---

---

---

---

---

---

## Données en coupe transversale et données en séries temporelles

Données en séries temporelles

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

	2000	2001	2002	2003	2004	2005	2006
Emissions	555,6	557,6	548,7	551,0	552,3	555,1	541,3
PIB	1441,37	1468,10	1483,18	1499,31	1536,35	1565,48	1599,48

Source : EUROSTAT

21

---

---

---

---

---

---

---

---





## 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

28

## Synthèse à partir de l'exemple 1

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

	Emission de gaz à effet de serre en 2002 (en millions de tonnes de CO2)	PIB en 2003 (milliards d'euros)	Superficie (1000 km <sup>2</sup> )	Population (en millions)	Densité moyenne (en hab./km <sup>2</sup> )	Population active (en %)	PIB/habitant (en milliers d'euros)	Pollution par habitant (en kg CO2 par an)	pollution/PIB (en kg CO2 par euro)	Monnaie
Allemagne (1)	1 130,3	2 183,8	357 021	82,1	231	75	28,23	12,52	0,43	Mark
Autriche	46,4	223,923	83 858	8,3	59	67	28,90	11,24	0,42	Mark
Belgique	166,9	2 17,25	30 528	10,9	247	97	26,50	13,92	0,52	Mark
Bulgarie	21,2	17,768	110 911	7,7	68	71	2,31	9,25	4,01	Lev
Chypre	1,3	1 078,9	9 259	0,9	138	62	11,73	11,73	1,00	Mark
Danemark	23,8	188,503	43 094	5,5	128	72	34,27	13,41	0,39	Couronne danoise
Estonie	13,7	8 892,3	45 226	1,3	29	89	8,89	15,15	2,27	Couronne estonienne
Etats-Unis	6 513,9	10 623,1	9 528 937	282,3	31	79	32,09	22,81	0,71	Dollar
France	64,6	145,928	37 033	6,3	18	62	27,54	18,00	0,65	Mark
Grèce	23,9	124 814,4	114 247	13,9	29	77	20,05	8,61	0,43	Mark
Irlande	133,5	171,408	131 840	11,2	85	89	15,30	11,92	0,78	Mark
Israël	6,8	14 278,7	33 300	10,4	109	85	1,98	1,68	0,84	Sheqel
Italie	28,6	120 441,9	2 9833	4,4	28	80	31,83	15,80	0,49	Mark
Japon	1 141,1	1 126 203,7	377 930	129,4	137	65	25,54	9,61	0,38	Yen
Lettonie	1 239,1	373,559	37 283	12,7	338	79	22,32	18,49	0,83	Mark
Lituanie	1,8	1 072,1	62 669	0,4	39	85	1,84	4,74	2,52	Mark
Malaisie	21,0	38 497,1	335 000	3,4	20	87	4,85	8,14	1,67	Ringgit
Malte	0,4	4 421,4	215	0,4	1 209	35	11,05	7,85	0,70	Mark
Népal	11,7	24 824,4	1 468	18,4	259	81	6,91	13,30	1,92	Mark
Ordonne	34,6	101 823,8	82 031	38,1	459	82	6,03	10,09	2,21	Zloty
Ordonne	13,0	1 18 593,1	11 695	10,1	44	75	12,25	7,73	0,63	Mark
Ordonne	145,5	10 924,1	7 269	13,4	141	74	7,86	13,14	1,65	Couronne lituanienne
Ordonne	168,9	142 020,9	244 020	61,2	249	30	67,00	13,80	0,20	Mark
Ordonne	10,9	24 465,1	2 016	0,4	111	26	1,46	3,26	1,72	Couronne slovaque
Ordonne	12,8	26 220,1	2 003	0,4	32	49	12,81	9,91	0,77	Mark
Ordonne	17,8	14 245,1	4 000	0,4	11	84	14,36	8,10	0,56	Couronne slovaque
Ordonne	17,8	207 220,4	41 000	2,4	109	88	18,30	7,70	0,42	Mark
Ordonne	12 180,2	14 302,4	14 140 000	10,4	9	—	35,40	14,64	0,41	Mark

29

## Synthèse à partir de l'exemple 1

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

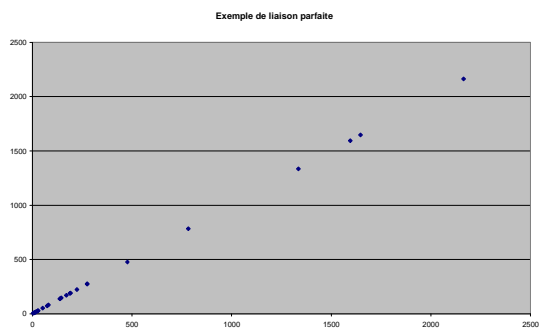
Lien entre population et PIB ?

Lien entre pollution et PIB ?

Lien entre pollution et densité de pollution ?

30

## Synthèse à partir de l'exemple 1 : liaison



---

---

---

---

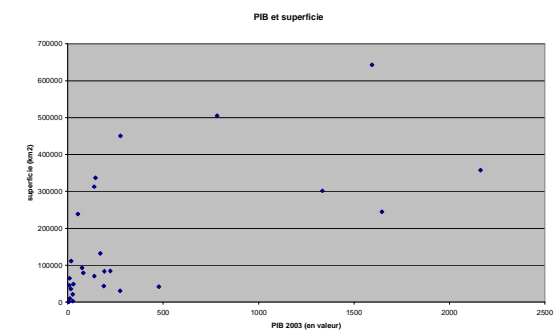
---

---

---

---

## Synthèse à partir de l'exemple 1 : liaison



---

---

---

---

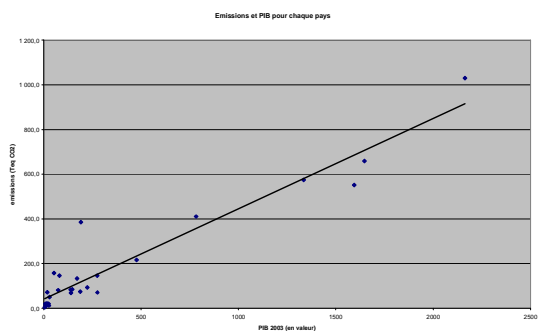
---

---

---

---

## Synthèse à partir de l'exemple 1 : liaison



---

---

---

---

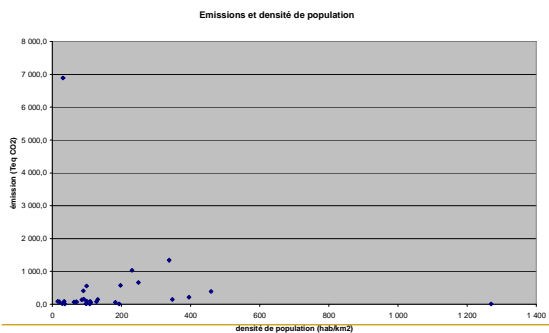
---

---

---

---

## Synthèse à partir de l'exemple 1 : liaison



34

---

---

---

---

---

---

---

---

## Chapitre 1 : tableaux et graphiques

35

---

---

---

---

---

---

---

---

## Plan

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

36

---

---

---

---

---

---

---

---

## Introduction : Lecture d'un tableau

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

	Cursus Licence	Cursus Master	Cursus Doctorat	Effectif total
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	84927	17960	2746	105633
Sciences humaines et sociales	135396	63463	14759	213618
Pluri-lettres-langues-sciences humaines	2505	3167	28	5700
Sciences fondamentales et applications	77420	65371	15899	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	19580	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).

37

## 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

38

## 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

39

## 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

40

---

---

---

---

---

---

---

---

## 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
Pepsi 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)

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

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

41

---

---

---

---

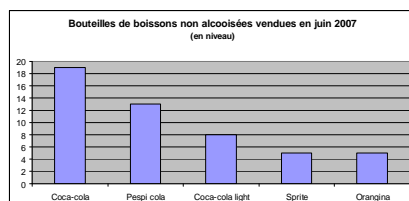
---

---

---

---

## 2. Données qualitatives : graphiques



42

---

---

---

---

---

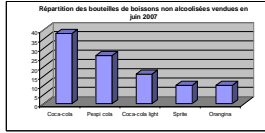
---

---

---

## 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  $0,38 \times 360 = 136,8^\circ$



43

## 2. Données qualitatives : tableaux pluri-dimensionnels

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

	Cursus			Fréquence totale
	Licence	Master	Doctorat	
	Fréquence	Fréquence	Fréquence	
Droit, sciences politiques	7,82	4,70	0,61	13,13
Sciences économiques, gestion (hors AES)	5,54	4,14	0,33	10,01
Administration économique et sociale (AES)	2,27	0,52	0,00	2,79
Lettres, sciences du langage, arts	4,89	1,73	0,51	7,11
Langues	6,16	1,25	0,20	7,61
Sciences humaines et sociales	9,93	4,65	1,08	15,66
Philosophie, lettres, langues, sciences humaines	0,18	0,23	0,00	0,42
Sciences fondamentales et applications	5,69	4,79	1,17	11,64
Sciences de la nature et de la vie	2,88	1,43	0,80	5,11
STAPS	1,87	0,45	0,04	2,36
Phyto-science	1,52	0,10	0,01	1,64
Médecine - Odontologie	4,07	7,52	0,08	11,66
Pharmacie	0,86	1,43	0,04	2,34
<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	6,52	7	0	13,52
<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).

44

## 2. Données qualitatives : tableaux pluri-dimensionnels

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

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

45

## 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	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	63,60	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	49,79	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>

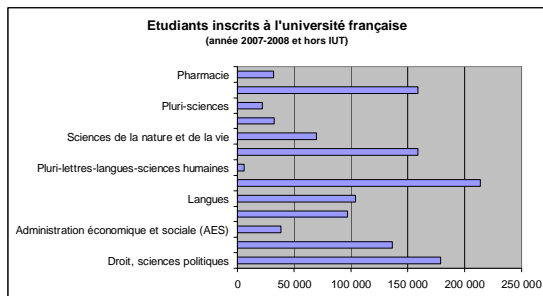
(/): absence de résultat due à la nature des choses.

Champ : France.

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

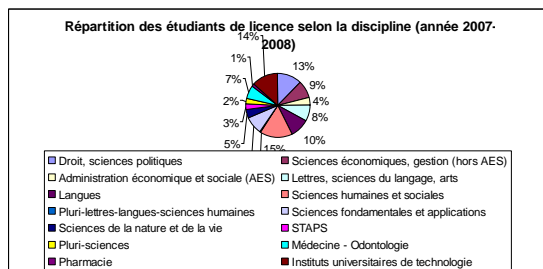
46

## 2. Données qualitatives : graphiques



47

## 2. Données qualitatives : graphiques



48

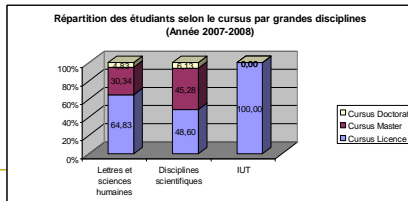
## 2. Données qualitatives : regroupements

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>	<b>62</b>	<b>33</b>	<b>5</b>	<b>100</b>

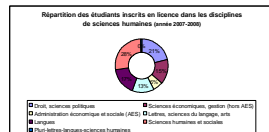
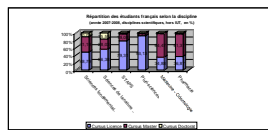
Champ : France.

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



49

## 2. Données qualitatives : regroupements



50

## 2. Données qualitatives : graphiques

Étudiants des universités par discipline

	2005-2006	2006-2007	2007-2008
<b>Total</b>	<b>1 421 719</b>	<b>1 359 177</b>	<b>1 363 750</b>
Chât. sciences politiques	175 833	173 265	170 225
Sciences économiques, gestion (hors AES)	154 706	154 729	155 474
Administration économique et sociale (AES)	44 451	41 369	38 002
Lettr. sciences du langage, arts	111 557	108 829	103 813
Sciences humaines et sociales	260 715	257 563	253 918
Plus lettres-langues-sciences humaines	4 247	6 576	8 700
Sciences fondamentales et appliquées	107 576	108 737	108 889
Sciences de la nature et de la vie	72 289	71 329	69 742
STAPS	41 519	40 481	39 159
Plus-science	21 817	21 181	22 301
Sciences technologiques	142 559	144 483	143 569
Pharmacie	28 624	31 293	31 871
<b>Total IUT</b>	<b>3 000 000</b>	<b>2 960 000</b>	<b>2 940 000</b>
<b>Total universités de technologie</b>	<b>139 889</b>	<b>133 269</b>	<b>133 263</b>



Champ : France.

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

51



### 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)	nombre (en milliers) (n)	amplitude (a)	densité (d=n/a)	effectifs corrigés n <sub>c</sub> = d * min(a)
0 - 19 ans	14 115	20	705,75	7057,5
20 - 29 ans	7 405	10	740,5	7405
30 - 39 ans	7 542	10	754,2	7542
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 - 89 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

55

---

---

---

---

---

---

---

---

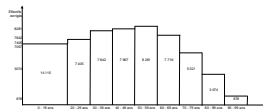
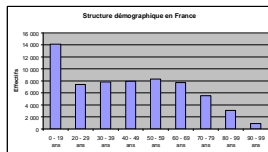
---

---

---

---

### 3. Données quantitatives : regroupements quantitatifs



56

---

---

---

---

---

---

---

---

---

---

---

---

### 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	8	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>

57

---

---

---

---

---

---

---

---

---

---

---

---

### 3. Données quantitatives : regroupements qualitatif

Regroupements par zone géographique

	nombre de pays	Fréquence	Emissions de gaz à effet de serre en 2002 (en millions de tonnes CO <sub>2</sub> )	Emissions de gaz à effet de serre en 2003 (en millions de tonnes CO <sub>2</sub> )	PIB en 2003 (Milliards d'euros)	PIB en 2003 (en %)	Superficie (km <sup>2</sup> )	Superficie (en %)	Population (en millions)	Population (en %)
Europe	15	55,56	33,92,97	69,33	7515,02	74,34	2948743	64,91	321,30	64,63
hors_Zone_E	12	44,44	17,43,95	33,67	2593,39	25,66	1532788	35,09	179,80	35,37
<b>Total</b>	<b>27</b>	<b>100</b>	<b>51,36,92</b>	<b>103,00</b>	<b>10108,41</b>	<b>100,00</b>	<b>4481531,00</b>	<b>100,00</b>	<b>501,10</b>	<b>100,00</b>

58

## Chapitre 2 : Méthodes numériques permettant de résumer une série

59

## 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

60

## Introduction

Deux étudiants peuvent avoir des moyennes identiques mais avec des dispersions 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.

61

---

---

---

---

---

---

---

---

## Statistiques résumant la tendance centrale : moyenne

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

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

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)

62

---

---

---

---

---

---

---

---

## 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)

63

---

---

---

---

---

---

---

---





## Statistiques résumant la tendance centrale : médiane

Distribution des notes pour le restaurant Y

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

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

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

$$Me = 3,85 \cdot 0,14 + 3 = 3,54$$

70

---

---

---

---

---

---

---

---

---

---

---

---

## 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 cumulée (%)	centre de classe (c)	$f_c$
[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	38,29
[600; 700[	95	45,24	83,33	650	284,05
[700; 800[	30	14,29	97,62	750	107,14
[800; 1000]	5	2,38	100	900	21,43
Total	210	100,00			
Moyenne					588,81

Source : B. PY (2009)

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

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

71

---

---

---

---

---

---

---

---

---

---

---

---

## 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

72

---

---

---

---

---

---

---

---

---

---

---

---

## 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

73

---

---

---

---

---

---

---

---

## 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

Soit un PIB =  $\frac{762,929 + 1335,3537}{2} = 1059,14$

74

---

---

---

---

---

---

---

---

## 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

75

---

---

---

---

---

---

---

---

## 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

76

---

---

---

---

---

---

---

---

---

---

---

---

## Statistiques résumant la tendance centrale : mode

Exemple 2 : variables quantitatives continues

âge (x)	nombre (en milliers) (n)	amplitude (a)	densité (d=n/a)	effectifs corrigés $h_i \cdot a_i =$
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 - 89 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

77

---

---

---

---

---

---

---

---

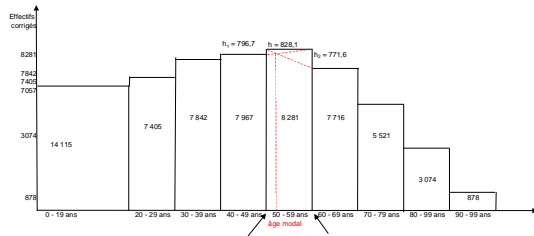
---

---

---

---

## Statistiques résumant la tendance centrale : mode



$$Mo = \frac{(h - h_1) \cdot x_2 + (h - h_2) \cdot x_1}{(h - h_1) + (h - h_2)} \quad Mo = \frac{(828,1 - 796,7)60 + (828,1 - 771,6)50}{(828,1 - 796,7) + (828,1 - 771,6)} = 53,57$$

78

---

---

---

---

---

---

---

---

---

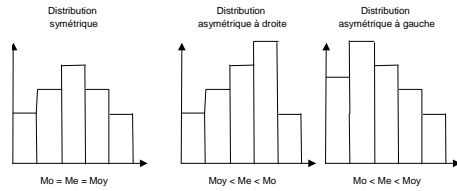
---

---

---

## Statistiques résumant la tendance centrale : discussion

Moyenne, mode et médiane et forme d'une distribution



79

---

---

---

---

---

---

---

---

## 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
Ouvriers	1000	10	1500	15
Cadres 1	3000	2	2000	11
Cadres 2	5000	1	2500	1
Total	9000	13	6000	17
Moyenne	615		1500	
Mode	1000		1500	

80

---

---

---

---

---

---

---

---

## 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.  
P<sub>2</sub> : valeurs extrêmes peuvent être très différentes des autres valeurs
- Intervalle interquartile ou interdécile : Q<sub>3</sub> – Q<sub>1</sub> ou D<sub>9</sub> – D<sub>1</sub>  
Délimitent la plage au sein de laquelle 50% ou 80% des valeurs sont regroupées  
Plus ces plages sont larges, plus les valeurs sont dispersées.  
P<sub>8</sub> : ne pas prendre en compte toutes les valeurs

81

---

---

---

---

---

---

---

---

## 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

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

82

---

---

---

---

---

---

---

---

---

---

## Statistiques résumant la dispersion

Notes des étudiants

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

83

---

---

---

---

---

---

---

---

---

---

## Statistiques résumant la dispersion : calculs

Distribution des notes pour le restaurant Y

Note	Effectif	$n_i$	$n_i(x_i - \bar{x})^2$
1	2	2	18,00
2	6	12	19,87
3	10	30	6,72
4	13	52	0,42
5	19	96	25,46
Total	50	191	69,38
Moyenne (X)		3,82	
variance		1,39	
écart-type		1,18	
coeff. Var.		0,31	

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

PIB pour 20 pays

Pays	PIB en 2003 (Milliards d'euros)	(x-X)	(x-X) <sup>2</sup>
Malte	4,2114	-789,89	623821,80
Estonie	8,6926	-788,82	614192,82
Lettonie	9,9778	-784,33	615174,82
Chypre	11,786	-789,92	613443,20
Lituanie	16,4971	-777,81	604990,75
Bulgarie	17,9668	-776,84	603517,18
Slovenie	25,7359	-768,67	590704,01
Luxembourg	34,8343	-768,47	590552,77
Chorésie	29,4856	-764,62	584644,24
Roumanie	2,611	-741,70	550112,38
Hongrie	32,5796	-719,73	518009,86
Republique tchèque	30,9241	-719,39	517417,46
Pologne	158,5821	-665,73	425277,26
Irlande	139,4419	-654,87	428850,41
Grèce	149,939	-648,37	420384,45
Israël	171,4998	-622,90	388002,63
Danemark	188,5003	-605,81	367003,71
Prusse	191,6438	-602,66	363204,87
Autriche	221,3021	-571,01	326048,21
Belgique	274,726	-519,58	269986,09
Suède	275,607	-519,05	269599,49
Suisse	287,7338	-506,85	256907,78
Pays-Bas	476,945	-317,36	100719,66
Espagne	782,929	-11,38	129,50
Italie	1338,3537	541,05	292739,79
France	1961,8311	800,51	640809,89
Royaume-Uni	1647,0556	829,75	678777,43
Allemagne (L)	2163,8	1369,49	1875006,67
Japon	3743,5796	2949,26	8699307,40
Etats-Unis	9809,8332	8889,22	79019200,51
Total de l'échantillon	23,92843	0,00	101999099,36

moyenne (X) 794,31  
variance 339970,00  
écart-type 1843,90  
coeff. Var. 2,32

84

---

---

---

---

---

---

---

---

---

---

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

Dépenses mensuelles en emplois à domicile

Dépense en euros	Effectifs	centre de classe (c <sub>i</sub> )	n <sub>i</sub> c <sub>i</sub>	n <sub>i</sub> (c <sub>i</sub> -X) <sup>2</sup>
[300; 400[	5	350	1750,00	309530,90
[400; 500[	60	450	27000,00	1328656,46
[500; 600[	15	550	8250,00	35735,54
[600; 700[	95	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)

85

## 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)

86

## Conclusion

Données macroéconomiques pour les pays de l'UE à 27

	Emissions de gaz à effet de serre en 2003 (en millions de tne CO <sub>2</sub> )	PIB en 2003 (Milliards d'euros)	Population (en millions)	Densité moyenne (en hab./km <sup>2</sup> )	PIB/habitant (en milliers d'euros)	Population par habitant (en tne CO <sub>2</sub> )	pollution/PIB (en kg de CO <sub>2</sub> par euro)
Allemagne	1 030,1	2183,3	82,3	231	26 29	12 62	0,48
Autriche	89,4	223 302,3	8,3	99	26 80	12 20	0,45
Belgique	1 48,3	274 220	10,6	347	25 92	13 50	0,53
Danemark	77,2	17 008	5,2	87	4 98	9 20	0,49
Espagne	3,2	11 288	4,0	108	11 19	9 20	0,79
Finlande	23,8	188 502,3	5,5	128	34 27	13 41	0,39
France	1 100	202 220	63,3	92	17 20	9 00	0,52
Grèce	19,7	8 602,6	1,3	29	6 69	18 10	2,77
Irlande	36,8	104 934	0,3	16	27 64	16 00	0,58
Italie	551,0	159 413	61,6	89	25 98	8 60	0,35
Pays-Bas	138,3	171 400,8	1,2	66	16 34	11 60	0,70
Portugal	80,4	78 426	10,1	109	7 86	7 60	1,08
République tchèque	89,4	138 443,3	4,4	69	17 89	10 60	0,49
Italie	674,1	1316 301,7	69,3	197	22 62	9 60	0,43
Malte	0,6	4 677,8	0,4	92	4 30	2 70	1,09
Espagne	21,0	16 497,1	4,4	69	4 85	6 10	1,27
Luxembourg	11,7	26 848	0,5	58	51 80	2 30	0,45
Malte	0,4	4 611,4	0,4	1 270	11 60	7 60	0,65
Pays-Bas	110,6	216 948	16,4	285	20 08	13 10	0,65
Pologne	189,1	152 653,8	38,1	209	6 00	10 00	0,60
Portugal	89,4	138 821	10,7	81	12 65	7 70	0,60
République tchèque	1 48,3	80 924,1	10,9	121	7 88	14 10	1,80
Roumanie	158,0	89 813	21,6	81	2 44	7 20	2,88
Slovaquie	68,0	107 656,6	5,0	269	47 00	10 00	0,42
Estonie	80,2	29 882,6	1,4	111	6 46	9 20	1,70
Autriche	19,8	26 778	8,0	69	12 91	9 60	0,77
Italie	19,2	228 857	8,1	89	30 49	7 70	0,28
Union européenne	2 179,2	10 086,4	497,1	113	26 85	16 40	0,61

Source : Eurostat (2004), p. 109, de 1991

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

87

## Conclusion

Données résumées pour les 27 pays de l'UE

	Emissions de gaz à effet de serre en 2003 (en millions de t eq 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 t eq CO2)	pollution/PIB (en kg eq CO2 par euro)
Moyenne	191,55	374,99	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	95,98	15,30	9,89	0,60
Q1	24,00	25,73	3,40	69,00	6,69	7,88	0,45
Q2	83,00	139,44	9,10	95,98	15,30	9,89	0,60
Q3	218,30	275,86	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
Écart-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

88

## Chapitre 3

### Indices et taux de croissance

89

## Plan

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

90

## 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	
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

91

---

---

---

---

---

---

---

---

---

---

---

---

## 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 %)

Villes	2000	2008
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>

92

---

---

---

---

---

---

---

---

---

---

---

---

## Comparaisons de données : Ecarts 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

93

---

---

---

---

---

---

---

---

---

---

---

---

## Comparaisons de données : Ratio

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

CA et CA/population					
	CA (en millions d'euros)	Rang	Population (en milliers)	CA/population (en millions d'euros)	Rang
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>	

94

## 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 =  $V_1 - V_0$

Variation relative = taux de croissance

$$= ((V_1 - V_0) / V_0) * 100$$

$$= (V_1 / V_0 - 1) * 100$$

95

## Mesures de l'évolution

Villes	CA (en millions d'euros)		Evolutions	
	2000	2008	Ecart absolu (en millions d'euros)	écart relatif (en %)
Brest	10000	11000	1000	10
Caen	8000	9000	1000	12.5
Nantes	20000	27000	7000	35
Rennes	15000	18000	3000	20
<b>Total</b>	<b>53000</b>	<b>65000</b>	<b>12000</b>	<b>22.64</b>

96

## 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

97

---

---

---

---

---

---

---

---

---

---

## 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%)

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

$$g_Q \approx g_x - g_y$$

98

---

---

---

---

---

---

---

---

---

---

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

	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Produit	1315,26	1387,87	1441,97	1497,17	1548,56	1594,91	1650,19	1726,07	1807,86	1892,24
Taux de croissance		5,41	3,87	3,67	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

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

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

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

...

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

99

---

---

---

---

---

---

---

---

---

---

## 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%

100

---

---

---

---

---

---

---

---

---

---

## 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 ?

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

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

Contribution à la croissance du CA de Machin

Villes	CA en millions d'euros		Parts	Taux de croissance	Contribution
	2000	2008	2000		
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>

101

---

---

---

---

---

---

---

---

---

---

## 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 €

102

---

---

---

---

---

---

---

---

---

---

## 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  $I_{t/0} = G/G_0$

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

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

103

## Les indices élémentaires

Produit intérieur brut aux prix de marché (en valeur)										
	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Niveau	1315,25	1367,97	1441,37	1487,17	1548,55	1594,81	1620,19	1725,07	1807,45	1892,24
Taux de croissance	4,01	3,97	5,07	3,37	3,83	2,89	1,53	6,07	4,72	4,69
Indice (base 100 en 1998)	100	104,01	109,59	113,83	117,74	121,25	122,42	131,23	137,42	143,97
Indice (base 100 en 2002)	84,53	89,34	93,08	99,69	100,00	102,29	107,21	111,46	116,72	122,19
Indice chaîné	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
base 100 en t-1		104,01	105,59	103,87	103,43	102,29	104,10	103,89	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

104

## Les indices élémentaires : propriétés

### Circularité

Base 1:  $I_{12/10} = I_{12/11} * I_{11/10}$

Base 100:  $I_{12/10} = I_{12/11} * I_{11/10} * 100$

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

$I_{11/10} = 1 / I_{10/11}$

105

## Les indices synthétiques

Comment synthétiser l'évolution simultanée de plusieurs variables.

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

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

106

## Les indices synthétiques

Indice de valeur :

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

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

107

## Les indices synthétiques : Indice de Laspeyres

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

⇒ Seuls les prix évoluent

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

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

108

## 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

109

---

---

---

---

---

---

---

---

## 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

110

---

---

---

---

---

---

---

---

## Chapitre 4

Corrélation et liaisons entre des variables

111

---

---

---

---

---

---

---

---

## 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 ?

112

---

---

---

---

---

---

---

---

## 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 ?

113

---

---

---

---

---

---

---

---

## 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)

114

---

---

---

---

---

---

---

---

## Données quantitatives : nuages de points

CA et spots publicitaires pour le magasin Truc

Semaines	Nombre 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 ?

115

---

---

---

---

---

---

---

---

---

---

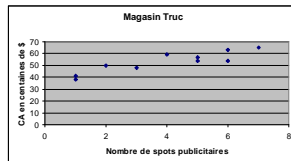
---

---

## 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



116

---

---

---

---

---

---

---

---

---

---

---

---

## 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 varient 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.

117

---

---

---

---

---

---

---

---

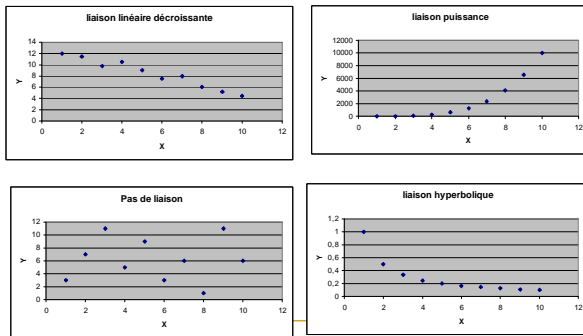
---

---

---

---

## Nuages de points : formes de liaison



118

---

---

---

---

---

---

---

---

---

---

---

---

## 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

119

---

---

---

---

---

---

---

---

---

---

---

---

## Covariance : calculs

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

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

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
Moyenne	4	52,9	227,1

Covariance = 227,1 - 4 \* 52,9 = 15,5

120

---

---

---

---

---

---

---

---

---

---

---

---

## 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.

121

---

---

---

---

---

---

---

---

## Coefficient de corrélation linéaire

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

Coefficient de corrélation linéaire

$$r = \frac{COV(X,Y)}{\sigma_x \sigma_y} \quad r = \frac{15,5}{2,049 * 8,37} = 0,903$$

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

122

---

---

---

---

---

---

---

---

## 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

123

---

---

---

---

---

---

---

---

## Régression linéaire

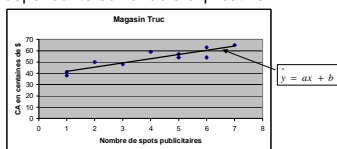
On cherche donc à estimer la droite qui s'ajuste le mieux au nuage de point

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



124

---

---

---

---

---

---

---

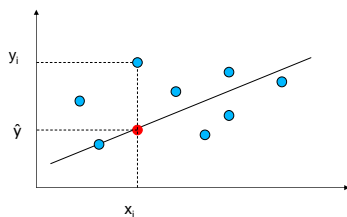
---

---

---

## Régression linéaire

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



125

---

---

---

---

---

---

---

---

---

---

## 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>
<b>Moyenne</b>	<b>4</b>	<b>52,9</b>	<b>227,1</b>	

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

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

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

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

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

126

---

---

---

---

---

---

---

---

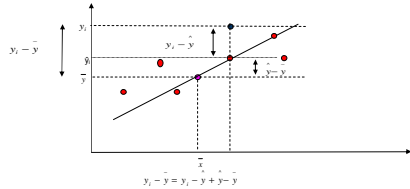
---

---

## 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  $(y_i - \bar{y})^2 = (\hat{y}_i - \bar{y})^2 + (y_i - \hat{y}_i)^2 \Rightarrow SCT = SCE + SCR$

127

---

---

---

---

---

---

---

---

---

---

## 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)	$\hat{Y}$	$(Y - \hat{Y})$	$(Y - m_0)^2$	$(\hat{Y} - m_0)^2$	$(Y - \hat{Y})^2$
1	2	50	45,52	4,48	9,41	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
Total	40	529			700,90	572,02	128,88
Moyenne	4	52,9			SCT	SCE	SCR

$a = \frac{15,54}{2} = 3,69$        $R^2 = \frac{SCE}{SCT}$   
 $b = 52,9 - 3,69 \cdot 4 = 38,14$        $R^2 = \frac{572,02}{700,9}$   
 $\hat{y} = 3,69x + 38,14$        $R^2 = 81,61$   
 $SCT = 572,02 + 128,88 = 700,90$        $R^2 = 81,61$

128

---

---

---

---

---

---

---

---

---

---

## 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

129

---

---

---

---

---

---

---

---

---

---