Méthodes quantitatives des sciences sociales

3. Tendance centrale et dispersion: deux points de vue complémentaires sur la réalité statistique

Sciences Po Saint-Germain-en-Laye, 1ère année

2016-2017

Introduction

- XIXème siècle: Adolphe Quételet (1796-1874) développe la théorie de « l'homme moyen » (dans *Sur l'homme et le développement de ses facultés, essai d'une physique sociale*, 1835, et *Sur l'appréciation des documents statistiques, et en particulier sur l'application des moyennes*, 1844). Les variations entre humains sont selon lui symétriques autour d'une valeur centrale, la **moyenne**. Elles suivent une **distribution normale**.
- 2009: dans le rapport Stiglitz-Sen-Fitoussi, les auteurs critiquent les usages de la moyenne, en regrettant qu'elle occulte une (grande) partie des phénomènes étudiés et, plus encore, de l'**expérience vécue** d'une partie de la population.

Introduction

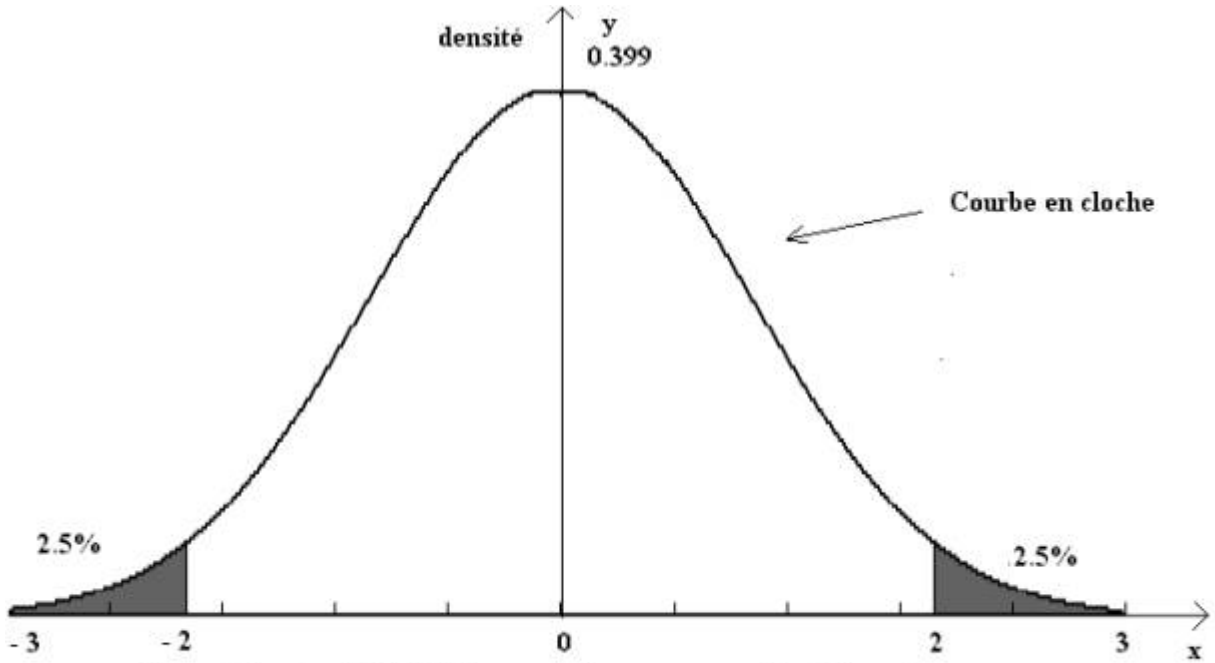
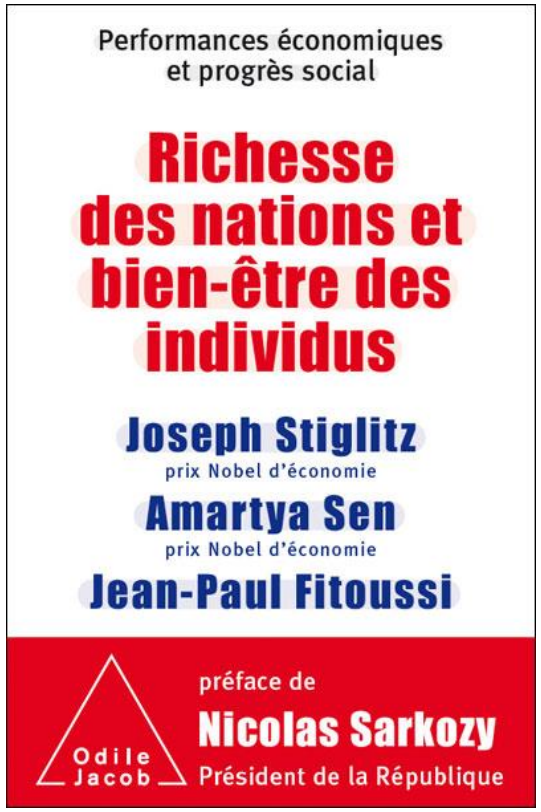


Figure 9.4 : densité de la loi normale de moyenne nulle et de variance un (courbe en cloche)

Introduction



Introduction

« Dans le monde entier, les citoyens pensent qu'on leur ment, que les chiffres sont faux… Et ils ont quelques raisons d'être dans cet état d'esprit. Pendant des années, on a dit à des gens dont la vie devenait de plus en plus difficile que leur niveau de vie augmentait. Comment ne se sentiraient-ils pas trompés ? (…) L'individu moyen n'existe pas et l'accroissement des inégalités le détache encore plus de l'expérience réelle de la vie » (Nicolas Sarkozy, préface du rapport Stiglitz-Sen-Fitoussi *Richesse des nations et bien-être des individus*, Paris, Odile Jacob, 2009).

Bibliographie

- INSEE, *Trente ans de vie économique et sociale*, Paris, INSEE, 2014.
- Jean-Paul Fitoussi, Amartya Sen, Joseph Stiglitz, *Richesse des nations et bien-être des individus*, Paris, Odile Jacob, 2009.
- Jean-Marc Meunier, *Statistiques pour psychologues. Analyses descriptives*, Paris, Dunod, 2010.
- Thomas Piketty, *Le capital au vingt-et-unième siècle*, Paris, Le Seuil, 2013.

Webographie

- http://topincomes.parisschoolofeconomics.eu/#Database: [Le site de la Top Incomes Data Base]
- http://piketty.pse.ens.fr/fr/capital21c [Le site du livre de Thomas Piketty]

Introduction

- Une distinction centrale: **variables quantitatives** (numériques) et **variables qualitatives** (catégorisées).
- On s'intéressera aujourd'hui aux seules **variables quantitatives**. On s'appuie notamment sur leurs caractéristiques **ordinales** (les modalités sont ordonnées).
- De façon classique, on représentera les valeurs observées de la variable par ordre croissant: de la plus petite à la plus élevée.
- Une procédure statistique courante : situer une observation par rapport à la **distribution** d'une variable. Exemples: niveau de vie d'un ménage ; moyenne au bac. Proportion de ménages (ou individus) plus extrêmes.

⇒Le **raisonnement simple** sous-jacent aux procédures d'inférence statistique.

Plan de la séance

- 1. Les indicateurs de tendance centrale: intérêt et limites
- 2. Les indicateurs de dispersion
- 3. De la dispersion aux inégalités

1. Les indicateurs de tendance centrale: portée et limites

- **Indicateurs de tendance centrale** ou **de position**
- **Idée générale**: on veut résumer une distribution observée par une valeur « centrale » qui la *représente bien*. Ou: *s'approche le plus de l'ensemble de la distribution*. Exemple: les notes d'un élève au Bac.
- **Première réponse** (qui est valable pour tous les types de variables, numériques ou non): on prend la valeur la plus fréquente parmi les valeurs observées. Exemple: l'élève a eu deux fois 11, et une seule fois chacune de ses autres notes. Cette valeur s'appelle le **mode**. Arbitraire.

1. Les indicateurs de tendance centrale: portée et limites

- Autre réponse: on calcule la **moyenne arithmétique simple** de l'ensemble des notes (poids implicite de chaque note: 1).
- On peut calculer plutôt une **moyenne pondérée**. Exemple: les « coefficients » des différentes matières sont les pondérations retenues. C'est la note qui aurait été obtenue à chaque épreuve si la performance générale était la même, et si l'élève avait réussi de la même façon chaque épreuve.
- Idée de **centre de gravité** de la distribution.

1. Les indicateurs de tendance centrale: portée et limites

- *I* ensemble d'individus. À l'élément *i* de *I* on associe sa valeur x^i
- La variable est notée x^I et l'effectif n_I. $\sum n_i = n$
- Fréquence

$$f_i = n_i / n$$

1. Les indicateurs de tendance centrale: portée et limites

- Total $t = \sum n_i \, x^i$
- Moyenne $Moy \, x^I = \bar{x} = \frac{t}{n} = \sum f_i \, x^i$
- Par définition, une **variable centrée** $(x^i - \bar{x})$ est de moyenne nulle

$$\sum n_i \, (x^i - \bar{x}) = 0 \; ; Moy \, (ax^I + b) = a \, Moy \, x^I + b$$

1. Les indicateurs de tendance centrale: portée et limites

Usage **très répandu** de la moyenne. La moyenne n'est pourtant pas toujours un si « bon » indice de tendance centrale.

Soit un pays à la population stable où le revenu annuel moyen est de 30000 dollars. Les 1% les plus riches reçoivent 20% du revenu total. En 5 ans, les revenus de ceux-ci (« le 1% ») augmentent tous de $+50\%$, alors que tous les autres revenus restent stables.

Le total des revenus augmente donc de $+10\%$ sans variation de population: le revenu annuel moyen est donc désormais de $30000+3000=33000$ dollars.

Pourtant la situation de 99% de la population n'a pas changé ! Un exemple caricatural ? Voire…

1. Les indicateurs de tendance centrale: portée et limites

- On a alors recours à la notion de **médiane**, qui est tout aussi intuitive. Notion de **milieu** (d'une distribution ordonnée).
- Si on ordonne une distribution de salaires, de revenus, de chiffre d'affaires…, la médiane est **la valeur qui partage cette distribution en deux parties égales**.
- Ainsi, pour une distribution de salaires, la médiane est le salaire au-dessous duquel se situent 50 % des salaires. C'est de manière équivalente le salaire au-dessus duquel se situent 50 % des salaires. (Source: INSEE.).

1. Les indicateurs de tendance centrale: portée et limites

- Dans l'exemple de tout à l'heure, la médiane est restée invariante, contrairement à la moyenne.
- **Autre exemple**: la négociation salariale dans une entreprise. Si seuls les hauts cadres ont vu leurs salaires fortement progresser l'an passé alors que tous les autres stagnaient, la moyenne peut avoir beaucoup augmenté. Le chef d'entreprise pourra communiquer sur cette hausse. Les représentants des salariés sont en droit d'estimer que cela cache un processus inégalitaire. Ils ont plutôt intérêt à raisonner en termes de **médiane** (s'ils veulent représenter l'ensemble des salariés) ou comparer les salaires moyens de sous-groupes.
- La médiane est moins **sensible aux variations extrêmes** et rend mieux compte de la distribution dans sa globalité (y compris sa *forme)*.

Illustration dans l'actualité récente en France (DARES)

- **Salaire moyen** d'un salarié à temps plein: 2128 euros en 2011.
- **Salaire médian à la même date**: 1712 euros nets par mois.
- La moyenne est appliquée à de nombreuses variables: espérance de vie, espérance de scolarité, nombre moyen d'élèves par classe, de pièces par logement, etc.
- La médiane est utilisée pour calculer le taux de pauvreté: proportion de ménages qui reçoivent moins de 60% du niveau de vie médian (« seuil de pauvreté »).

Illustration dans l'actualité récente en France (INSEE)

- Le revenu disponible d'un ménage comprend les revenus d'activité, les revenus du patrimoine, les transferts en provenance d'autres ménages et les prestations sociales (y compris les pensions de retraite et les indemnités de chômage), nets des impôts directs. Quatre impôts directs sont généralement pris en compte : l'impôt sur le revenu, la taxe d'habitation et les contributions sociales généralisées (CSG) et contribution à la réduction de la dette sociale (CRDS).
- Le niveau de vie est égal au revenu disponible du ménage divisé par le nombre d'unités de consommation (uc). Le niveau de vie est donc le même pour tous les individus d'un même ménage.
  Les unités de consommation sont généralement calculées selon l'échelle d'équivalence dite de l'OCDE modifiée qui attribue 1 uc au premier adulte du ménage, 0,5 uc aux autres personnes de 14 ans ou plus et 0,3 uc aux enfants de moins de 14 ans.

Illustration dans l'actualité récente en France (INSEE)

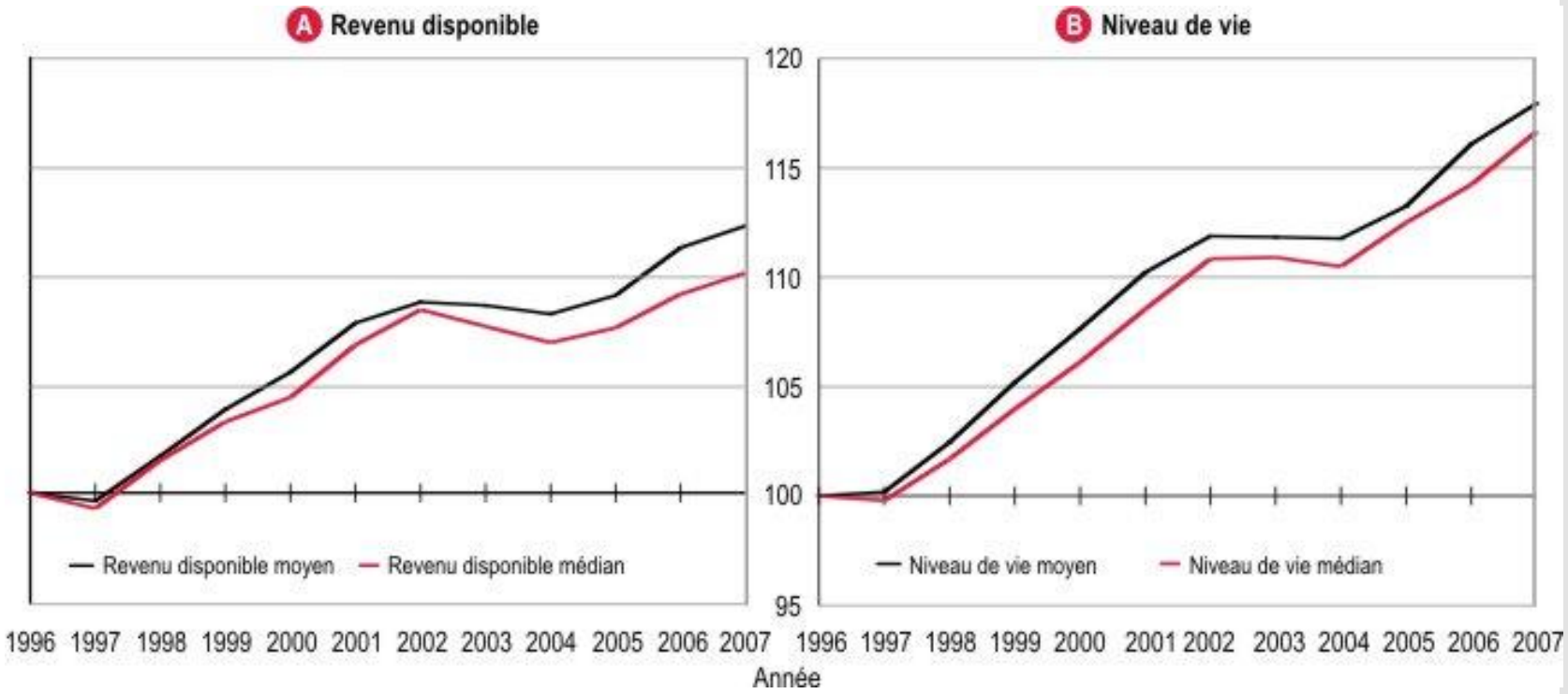
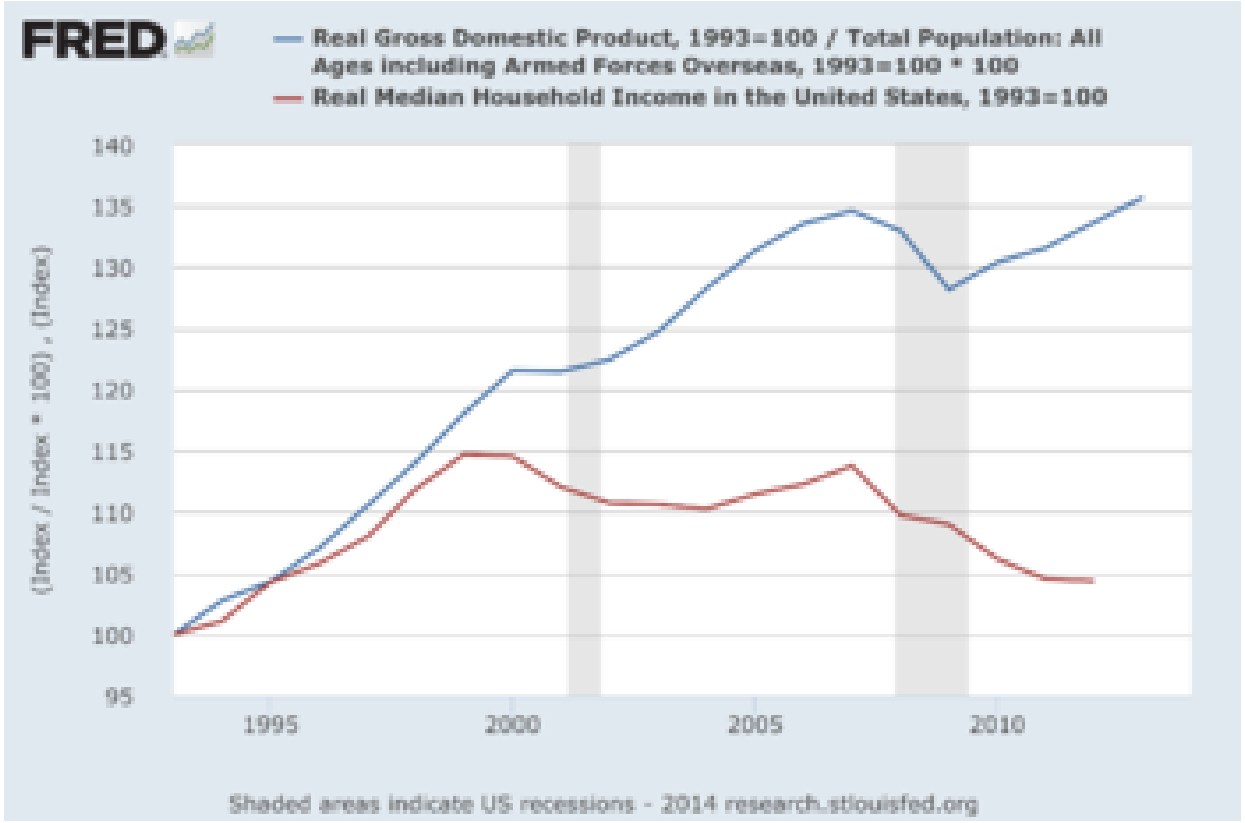


Illustration dans l'actualité récente aux Etats-Unis (PIB/hab, revenu median)



Commentaires

- En France, l'écart entre moyenne et médiane des revenus reste limité (jusqu'en 2007).
- Aux Etats-Unis, il semble plus net, et les évolutions semblent clairement divergentes, révélant un phénomène de fond (cf. Joseph Stiglitz, *Le prix de l'inégalité*, 2010). Attention, une autre différence dans les données présentées accentue l'écart: production / revenu.
- Une dynamique des revenus tirée par les hauts, voire très hauts, revenus. Les 1% de « Occupy Wall Street ». Pendant ce temps, les bas revenus sont maintenus bas (mondialisation, concurrence, réformes structurelles du marché du travail, etc.).
- Les très hauts revenus dépendent le plus des évolutions des marchés financiers, y compris du marché immobilier. Un phénomène spéculatif ? Notion de « bulle financière ».
- Des questions socio-économiques: une dynamique saine ou pathologique ? Conséquences de cette dynamique.

Conclusion du 1.

- Nécessité de combiner moyenne et médiane, et d'analyser d'éventuelles divergences entre les deux.

2. Les indicateurs de dispersion

- La tendance centrale ou position cache la **répartition** des valeurs.
- Idée générale, intuitive: selon que les notes à un examen apparaissent très éloignées, ou au contraire très rapprochées les unes des autres on les déclarera « très dispersées » ou « peu dispersées »… On dira que la classe est plus ou moins « homogène ».
- Dans certains cas, on peut comparer les dispersions de deux protocoles sans avoir besoin de formaliser davantage. Ex: (+14;+38;-2;-16;+22) ; (+19;+43;+3;-11;+27) ; (+26;+38;+18;+11;+30).
- Sinon, on va chercher un indice de dispersion (positif ou nul). Quel indice ?

2. Les indicateurs de dispersion

- Première idée: écarts à la moyenne. Non: la somme des écarts des observations à la moyenne est nulle.

$$\sum n_i \, (x^i - \bar{x}) = 0$$

(caractérisation barycentrique de la moyenne)

2. Les indicateurs de dispersion

- L'écart absolu moyen, *Eam*, est la moyenne des écarts absolus à la valeur moyenne.

$$Eam = \frac{\sum n_i |x_i - \bar{x}|}{n}$$

- *Positivité ; intermédiarité (entre valeurs extrêmes) ; si tous les écarts sont égaux, leur valeur commune est égale à l'eam.*

2. Les indicateurs de dispersion

- La variance est la moyenne pondérée des carrés des écarts à la moyenne.

$$Var\ x^I = \sum f_i\ (x^i - \bar{x})^2$$

- L'écart-type est la racine carrée de la variance.

$$Ety\ x^I = \sqrt{Var\ x^I}$$

2. Les indicateurs de dispersion

- Propriétés de la variance

$$Var\ x^I = \sum \frac{n_i\ (x_i)^2}{n} - \bar{x}^2$$

- La variance est la moyenne des carrés bruts diminuée du carré de la moyenne. Var = Moy car – car Moy.
- L'écart-type a les mêmes propriétés que l'eam. C'est un écart quadratique moyen.
- On a ainsi décrit **la moyenne de la dispersion autour de la moyenne**.
- Variance (et écart-type) sont donc associés à la notion de moyenne.
- Les distributions théoriques (comme la distribution normale) sont caractérisées par la moyenne et l'écart-type.
- Variable centrée-réduite: moyenne nulle et écart-type unitaire.

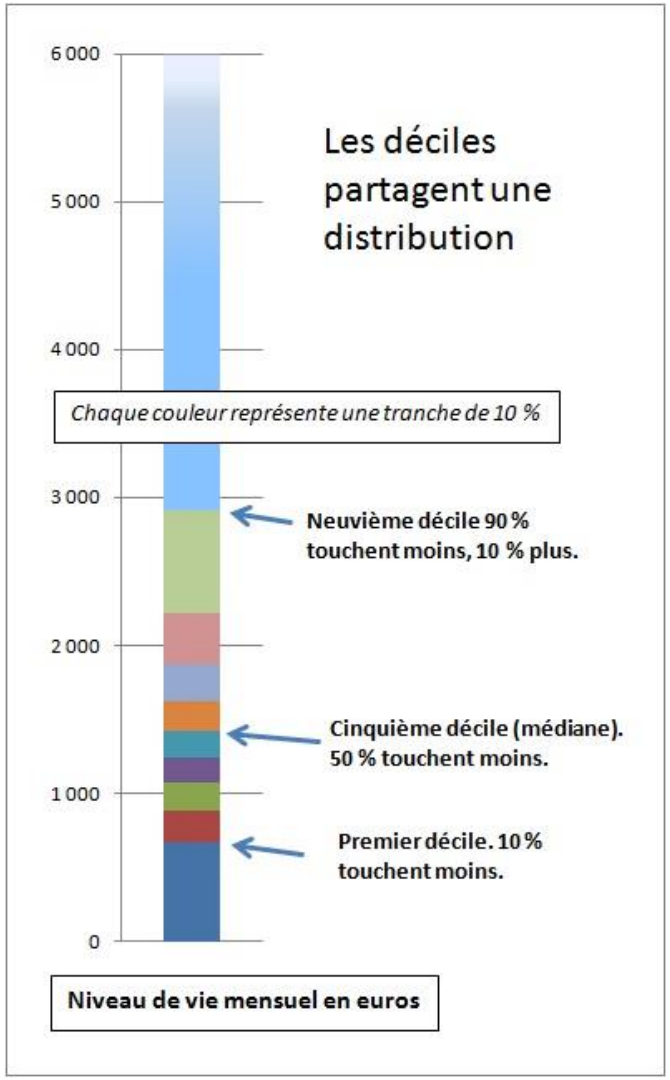
2. Les indicateurs de dispersion

Si on ordonne une distribution de salaires, de revenus, de chiffre d'affaires..., les quartiles sont les valeurs qui partagent cette distribution en quatre parties égales.
Ainsi, pour une distribution de salaires :
- le premier quartile (noté généralement Q1) est le salaire au-dessous duquel se situent 25 % des salaires ;
- le deuxième quartile est le salaire au-dessous duquel se situent 50 % des salaires ; c'est la médiane ;
- le troisième quartile (noté généralement Q3) est le salaire au-dessous duquel se situent 75 % des salaires.

Le premier quartile est, de manière équivalente, le salaire au-dessus duquel se situent 75 % des salaires ; le deuxième quartile est le salaire au-dessus duquel se situent 50 % des salaires, et le troisième quartile le salaire au-dessus duquel se situent 25 % des salaires.

(Source: INSEE).

2. Les indicateurs de dispersion



2. Les indicateurs de dispersion

- La généralisation de la notion de médiane: les quantiles. Quartiles, quintiles, etc. Ce sont les valeurs qui coupent la distribution ordonnée selon les grandeurs croissantes en n sous-populations.
- On va ensuite étudier les **relations entre quantiles**. On privilégiera la relation entre quantiles extrêmes. Exemple : 1er et 9ème décile.
- **Ecart inter-quantile**: Q9 – Q1, etc.
- **Rapport inter-quantile**: Q9 / Q1
- Ce sont là deux façons différentes de mesurer la dispersion, ou **indicateurs de dispersion**.
- Les écarts et rapports inter-quantiles sont en harmonie avec la médiane.
- Ils permettent de mesurer des **inégalités**.

Illustrations

- A la suite du rapport Stiglitz-Sen-Fitoussi de 2009, le Conseil d'analyse économique et le Conseil des sages allemand ont souhaité proposer de nouveaux indicateurs pour mesurer les performances économiques et le progrès social en France et en Allemagne.
- Le rapport inter-quintile dans la distribution des revenus (S80/S20) comme indicateur global d'inégalité de répartition.
- On peut appliquer le même raisonnement à toute variable numérique. Exemple: le nombre d'années de scolarité (Programme des nations unies pour le développement). Si le rapport inter-quantiles (extrêmes) augmente, on conclura à une hausse des inégalités ; les performances à des tests cognitifs (enquête PISA).
- Un choix qui implique une convention forte. Avec les quartiles, le « 1% » des plus riches disparaît dans la masse du premier quart. C'est le raisonnement de Thomas Piketty, qui multiplie les indicateurs à des niveaux fins (0,1%, 0,01%, etc.).

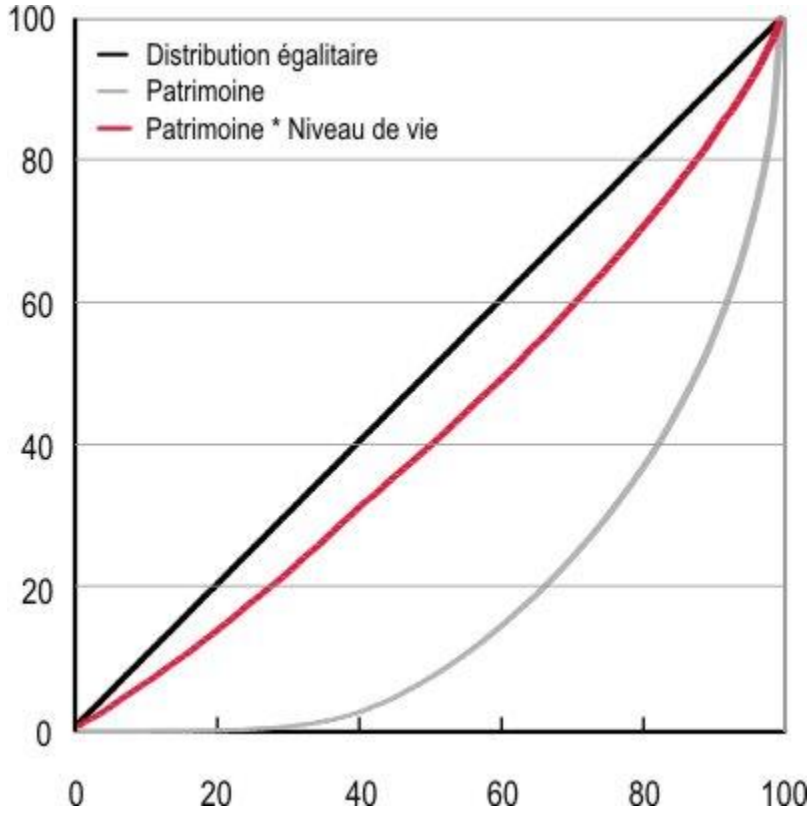
3. De la dispersion aux inégalités

- Dispersion des tailles, des poids, les âges, etc. On étudie les **différences** inter-individuelles, qui ne sont pas toujours interprétables directement en termes d'inégalités. Exemple de la taille: la notion d'inégalité ne semble pas très pertinente. Mais la taille peut générer des inégalités. Stigmatisation des petites tailles, etc. Idem pour le poids et l'âge.
- Une **inégalité** est une différence qui présente un caractère systématique et favorise (ou défavorise) les uns aux détriments (au profit) des autres.
- Notion à ne pas charger –au moins provisoirement- de caractère normatif (inégalité n'implique pas toujours injustice).
- Cela implique de s'intéresser non seulement à la distribution de la variable mais aussi à la grandeur mesurée et à sa répartition au sein de la population. Métaphore du gâteau: part du « gâteau » global (c'est-à-dire du total t défini plus haut). Cette notion n'a pas toujours de signification interprétable. Exemple des notes au bac. En revanche, pertinent pour certaines grandeurs économiques et sociales (revenus, patrimoines, années de vie, de scolarisation…).

3. De la dispersion aux inégalités

- On va représenter en abscisse la population en quantiles ordonnés, et en ordonnée le volume correspondant de la grandeur étudiée (ou total du revenu détenu par les n%). C'est la courbe de Lorenz: « la part de [revenu / patrimoine] détenue par les ménages lorsqu'on les classe par ordre de [revenu / patrimoine] croissant ».
- On va comparer cette courbe à une courbe correspondant à l'équi-répartition: tout le monde possède le même patrimoine / reçoit le même revenu. Situation d'égalité parfaite. Plus on s'en éloigne, plus l'on est confronté à une situation de **concentration** de la grandeur considérée (donc d'inégalité).
- Cela conduit à l'indice de Gini, qui est calculé à partir de la courbe de Lorenz comme rapport entre deux surfaces. Plus l'indice est élevé plus la situation est inégalitaire.

3. De la dispersion aux inégalités

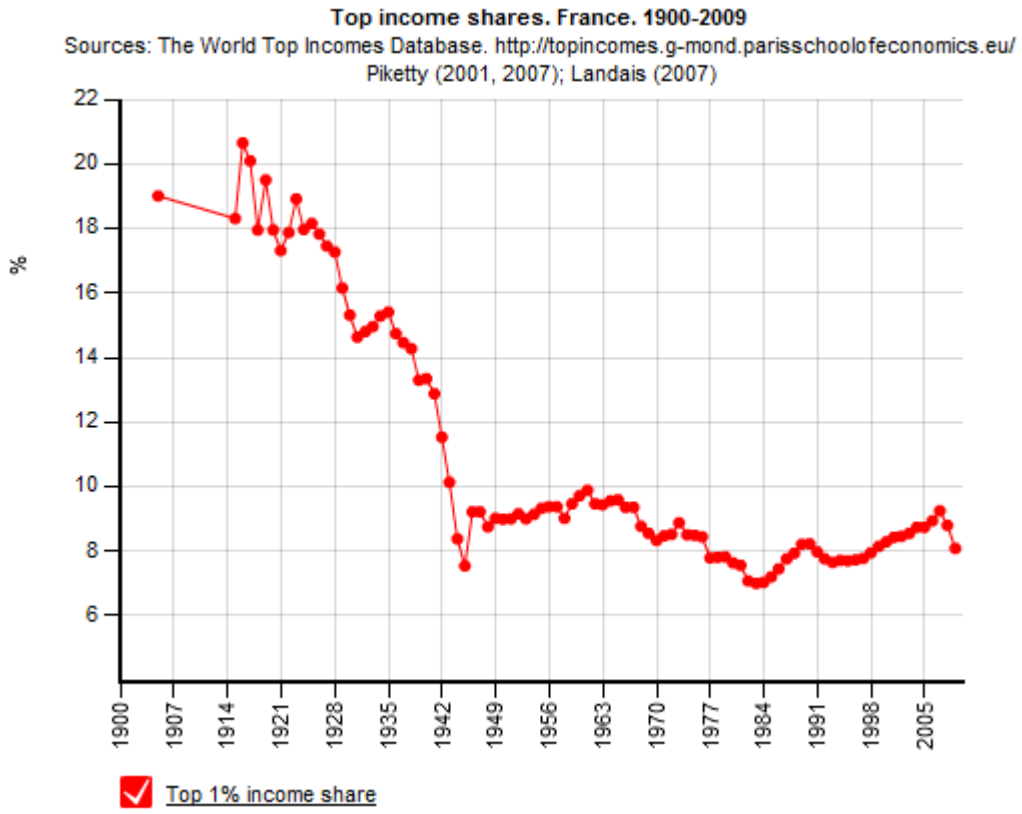


Lecture : la courbe de Lorenz (en gris) représente la part de patrimoine détenue par les ménages lorsqu'on les classe par ordre de patrimoine croissant. La pseudo-courbe de Lorenz (en rouge) représente la part du patrimoine détenue lorsque les ménages sont classés par ordre croissant du niveau de vie. Plus les courbes s'éloignent de la diagonale (en noir), plus la distribution est inégalitaire.
Champ : France métropolitaine, ménages ordinaires, montants recalés sur les données de la Comptabilité nationale. **Source** : Insee, enquête Patrimoine des ménages 2004 et Comptabilité nationale.

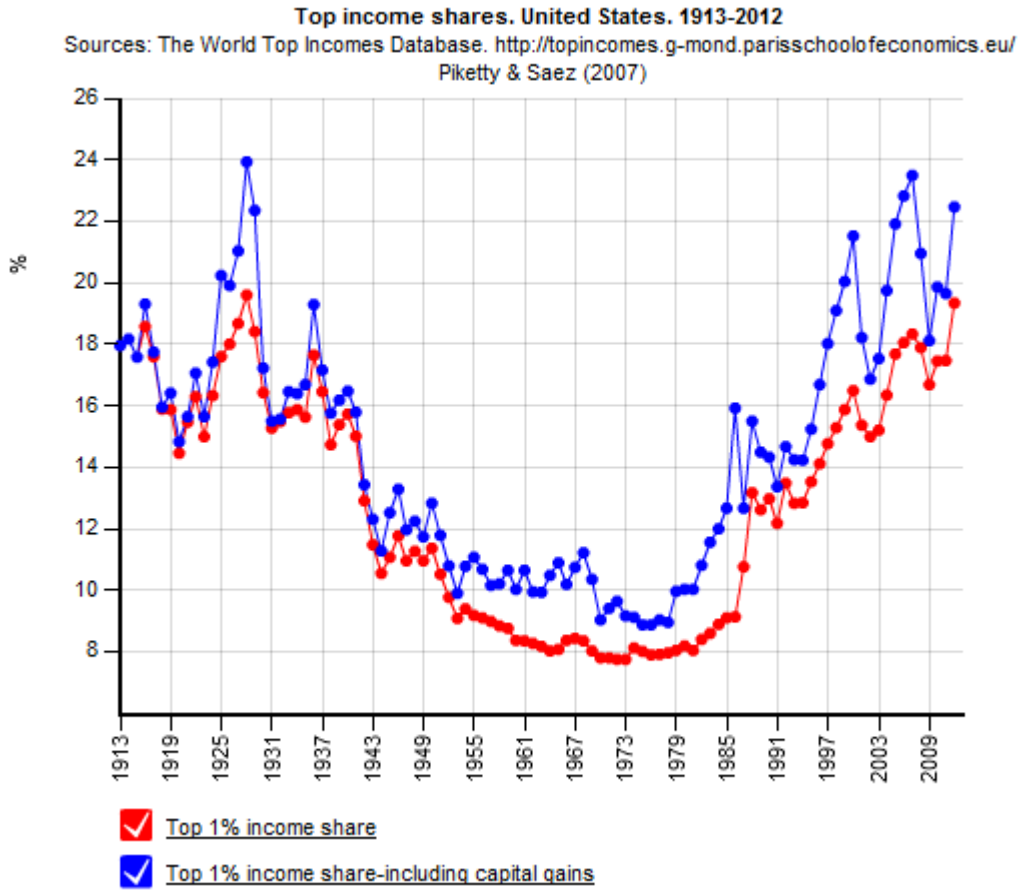
3. De la dispersion aux inégalités

- Indice de Gini des niveaux de vie en France: 0,31 en 2011.
- Une autre démarche consiste à mesurer la proportion de la grandeur totale détenue par le quantile d'intérêt (le 1% le plus riche par exemple) ou **contribution (d'un quantile) à une variable numérique**. En cas d'égalité parfaite: le 1% détient 1% du gâteau total. A l'opposé extrême, un individu détiendrait 100% de la richesse.
- C'est la démarche adoptée par Thomas Piketty et son équipe (cf. *Le capital au 21ème siècle, 2013*).

3. De la dispersion aux inégalités



3. De la dispersion aux inégalités



3. De la dispersion aux inégalités

- Le « débat Piketty »: un débat transnational.
- Constitution de **séries longues** d'inégalités dans plusieurs pays. Retour de l'histoire sérielle.
- La critique du *Financial Times:* un problème de sources (registres et enquêtes).
- L'enjeu des sources fiscales: la partie émergée d'un iceberg ? Cf. Pierre Lascoumes, *Sociologie des élites délinquantes. De la criminalité en col blanc à la corruption politique*, 2014 ; Gabriel Zucman, *La richesse cachée des nations. Enquête sur les paradis fiscaux*, Paris, 2013.
- *Quelle théorie ? r > g ?* La croissance et la rente: les revenus du capital en question.

3. De la dispersion aux inégalités

- Statistiques, mesures de la performance et conceptions de la justice.
- La performance **moyenne** dépend-elle du niveau d'**inégalités** ? Si oui, dans quel sens ?
- Si non, l'égalité est-elle souhaitable quel que soit le niveau moyen ?
- Empiriquement, les pays les plus riches sont *plutôt* plus égalitaires.
- OCDE, FMI: lien entre montée des inégalités et ralentissement de la croissance.
- Inégalité et financiarisation: le rôle des « bulles spéculatives ».

Conclusions

- Deux points de vue complémentaires sur une variable numérique: le « niveau global » ; les variations inter-individuelles. Nécessité de toujours articuler les deux.
- Deux logiques statistiques: moyenne-variance ; médiane-relations entre quantiles.
- Enjeu de la représentation graphique (*data visualisation*).
- La notion d'inégalité prolonge celle de dispersion, qui est plus générale.
- L'enjeu de l'interprétation des inégalités et de leur évolution.
- Inégalités et justice, une question encore très ouverte.