La Communication Autour de l'Incertitude des Statistiques Publiques

Ronan Le Saout (CREST-ENSAI)

JMS, Juin 2018

Motivations

- En tant que statisticien-économiste, il est souvent surprenant de constater que la précision des statistiques ne soit pas toujours publiée.

"Je dois reconnaître avoir souvent été mal à l'aise à me dire que les statisticiens français faisaient des efforts peut-être insuffisants pour renseigner sur la précision de leurs résultats." (Malinvaud 1988)
"Une statistique n'a pas de valeur sans son écart-type." (Dell *et al.* 2002)

- L'incertitude ne se limite pas à la précision, mais le constat reste le même.

"Si les statisticiens ne manquent pas, du fait même de la difficulté de leur métier, d'être attentifs à la qualité de l'information qu'ils produisent, il n'est pourtant pas usuel en France qu'ils examinent "publiquement" les risques d'erreurs ou de biais, ou évaluent l'imprécision et les révisions de leurs estimations. Le plus souvent, ces débats très techniques ne dépassent pas un cénacle de spécialistes." (Domergue 1995)

- Une question qui n'est pas nouvelle (Armatte 2003 citant Thionet 1954)

Motivations

- Peu de connaissance sur les raisons invoquées aux choix de communication, malgré un constat d'une évolution historique vers une plus grande transparence ;
- Manski (2015) :
  - Trois composantes à l'incertitude (permanente, transitoire et conceptuelle) ;
  - Incertitude qui se rapproche de la notion de qualité ;
  - Manque de communication des instituts anglo-saxons autour de l'incertitude des statistiques.
- Volonté d'étudier le cas français, en dépassant le simple constat, à partir d'une revue bibliographique et d'entretiens :
  - De la distinction entre statistiques descriptives et études économiques ;
  - Evolution historique et raisons afférentes ;
  - Rôle des différences conceptuelles de production.

Qu'est ce que l'incertitude ?

- Trois composantes (Manski 2015) :
  - Permanente : erreurs d'échantillonnage et de traitements d'enquête (précision) mais également non liées à l'échantillonnage (modes de collecte, questions, enquêteurs...) ;
  - Transitoires : statistiques provisoires, semi-définitives ou définitives ;
  - Conceptuelles : la "boîte noire" du statisticien, méthodes de désaisonnalisation, de correction de la qualité des indices...

Les modes de communication

- Communication grand public : tableaux, commentaires et “4 pages”, vulgarisation scientifique ;
- Communication spécialisée : document de travail, publication académique, séminaire, méta-données ;
- Communication détaillée : accès aux bases de données et informations associées.

Plan de la présentation

- 1 - Des différences théoriques d'approche entre économétrie et théorie des sondages
- 2 - Quelques raisons pour communiquer ou non l'incertitude
- 3 - Des choix qui dépendent des statistiques produites : issues d'enquêtes, de sources exhaustives, de sources multiples, ou de modélisation
- 4 - Discussion autour des Big Data et de l'Open Data

1 Statistiques descriptives et études économiques

- L'INSEE, à la fois producteur et utilisateur des données statistiques ;
- Théories utilisées : théorie des sondages d'un coté, économétrie de l'autre ;
- Notions d'aléa différentes, liées au tirage ou au modèle ;
- Un estimateur identique de la moyenne : Hajek en théorie des sondages, régression pondérée en économétrie ;
- Des choix en termes de communication de la précision complètement différents :
    - Pour des statistiques descriptives, précision peu mentionnée dans les publications généralistes.
    - Pour une étude économique, précision communiquée (p-value...).

1 Statistiques descriptives et études économiques

- Différence d'objectifs entre la description et l'explication ;
- Différence créée par les cursus de formation ;
- D'un point de vue plus sociologique :
  - Un modèle est associé à une étude des comportements nécessairement bruités ;
  - Une statistique descriptive renvoie à une demande sociale d'exactitude, mesuré par une institution "indiscutable".

2 Quelques raisons pour communiquer ou non l'incertitude

- Une communication à deux étages ;
- Rôle croissant de l'évolution technologique ;
- Rôle moteur des institutions nationales et européennes ;
- Arbitrage entre transparence, pédagogie et défense de l'institution ;
- Une réflexion en devenir : le contenu des bases détaillées.

2.1 Une communication à deux étages

- Communication des méthodes effectuée auprès de spécialistes à travers des publications académiques ou des séminaires, complétée par un calcul à usage interne de la précision (et autres formes d'incertitude) ;
- Mais information non aisément accessible au grand public ;
- Fortes implications des unités de l'INSEE (UMS, puis DMS, et pôle PISE) sur le calcul de la précision, effectuée désormais pour presque toutes les enquêtes ;
- Les contraintes théoriques, si elles existent, n'expliquent en général pas l'absence de diffusion de la précision des statistiques ;
- Communication également différenciée pour l'incertitude transitoire et conceptuelle, avec une évolution historique vers plus de transparence.

2.1 Une communication à deux étages

- Mais bases de données n'incluant pas l'ensemble des millésimes diffusés ;
- Mais chiffres "non transformés" pas toujours diffusés ;
- Coût lié à rendre disponible ces informations complémentaires potentiellement important.

2.2 Rôle croissant de l'évolution technologique

- Enjeux computationnels mineurs quant à la possibilité de calculs de la précision et des autres formes d'incertitude ;
- Coûts de stockage réduits de manière drastique depuis les années 1970 ;
- Choix historiques et contraints associés à des chaînes de production spécifiques, dont la logique n'a pas forcément été modifiée malgré la baisse des coûts de stockage ;
- 2003 : diffusion gratuite de la majorité des statistiques produites, avec une contrainte l'absence de demandes complémentaires ;
- L'incertitude n'étant pas diffusée au préalable, il n'a pas été envisagé d'étendre la communication, pour des questions de coût.

2.3 Rôle moteur des institutions nationales et européennes

- Une notion d'incertitude qui ne se limite pas au calcul de la précision ;
- Rôle moteur depuis les années 2000 à la fois des institutions nationales (ASP, Comité du Label) et européenne (code de bonnes pratiques) :
  - Fiche qualité sur le site internet, en particulier pour les enquêtes entreprises (avec des indicateurs de précision) ;
  - Bouton "M" sur le site d'Eurostat.
- Renforcement de la notion de qualité (code des bonnes pratiques de la statistique européenne en 2005) ;
- Pas de demandes des utilisateurs relatives à l'incertitude au sein des instances du CNIS.

2.4 Arbitrage entre transparence, pédagogie et défense de l'institution

- Un sentiment diffus de doute sur la crédibilité des statistiques existe, difficile à expliquer, contrôler ou infléchir (Charpin 2010) ;
- Desrosières (2003) “comme c'est le cas pour de beaux monuments, l'exhibition de ces échafaudages de métadonnées peut, qu'on le veuille ou non, brouiller, sinon la beauté, du moins l'efficacité argumentative de l'évidence factuelle impliquée par la mise en avant d'un simple nombre, tout nu.” ;
- Pédagogie autour des notions statistiques délicate, mais une illusion de certitude associée à certains choix de diffusion (nombre de décimales) ;
- Pas de règles générales, "la communication gardant une part de mystère, tant sont subtils les canaux qui vont influencer les utilisateurs dans leur jugement" (Charpin 2010).

2.5 Une réflexion en devenir : le contenu des bases détaillées

- Avancées récentes sur l'accès aux données depuis les années 2000 : CASD, Comité du Secret, insee.fr, Centre Quetelet ;
- Processus qui s'inscrit dans un contexte d'une demande de reproductibilité des études économiques ;
- Réflexion encore à mener sur le contenu de ces bases détaillées pour permettre une juste prise en compte de l'incertitude pour des utilisateurs extérieurs ;
- Par exemple, si certaines bases de données précisent les observations imputées, ce n'est en rien systématique ;
- Faute de connaissance des utilisateurs, il ne faut donc pas attendre une demande extérieure sur une meilleure communication des incertitudes entourant chaque processus d'enquête.

3 Les différences conceptuelles de production

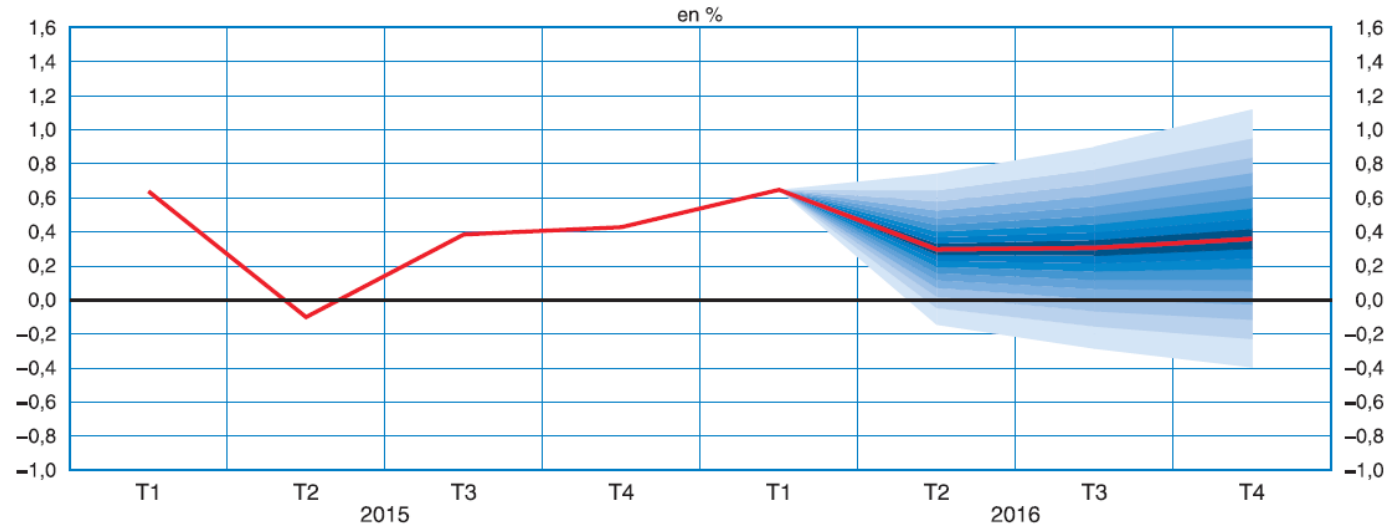
- Desrosières 2003 : les objets statistiques peuvent être vus « à la fois comme « réels » (ils existent antérieurement à leur mesure) et comme « construits à partir de conventions » (ils sont, d'une certaine manière, « créés » par ces conventions).
- Selon les sources utilisées, la perception de cette tension par les producteurs peut en être plus ou moins forte.
- Quatre types de production statistique distingués, les statistiques issues d'enquêtes, de sources administratives exhaustives, celles fruits de diverses sources et d'une construction économique, et enfin la prévision économique.

3 Les différences conceptuelles de production

- Pour les statistiques issues d'enquête, un questionnaire et un corpus mathématique (la théorie des sondages) ancrent le calcul dans la réalité.... mais nécessité de communiquer autour des différentes formes d'incertitudes, permanentes (précision, modes de collecte), provisoires notamment à travers les différents millésimes du recensement, et conceptuelles à travers les différences de mesures du taux de chômage ;
- Les données admistratives restent imprécises car elles ont pour usage premier une utilisation administrative et non statistique. Les règles administrives peuvent évoluer dans le temps, source d'incertitude conceptuelle (chiffres de Pôle Emploi, de la délinquance, des comptes des collectivités locales) ;
- Pour la comptabilité nationale (ou autres statistiques construites à partir de sources multiples), les conventions adoptées sont plus explicites, mais la notion de précision en devient aussi plus difficile voire impossible à définir. L'incertitude sera alors principalement abordée à travers les révisions et les concepts.

3 La conjoncture économique

- Les chiffres de prévision ne relèvent pas dans la majorité des pays des statistiques calculées par les instituts de statistique.
- L'INSEE a adopté en 2008 le graphique des risques diffusé par la Banque Centrale d'Angleterre, qui illustre graphiquement la précision et les intervalles de confiance des prévisions.
- Un des objectifs en termes de communication : les prévisions ne sont pas de même nature que les statistiques, elles ne s'appuient pas uniquement sur des données mais aussi sur une modélisation, plus ou moins explicite, des comportements économiques des agents.
- Exemple de la note de conjoncture de Juin 2016



Lecture : le graphique des risques retrace, autour de la prévision centrale (en trait rouge), 90 % des scénarios probables. La première bande, la plus

L'incertitude à l'ère des données massives et ouvertes

- Réappropriation de la notion de qualité des données par les chercheurs en sciences sociales ;
- Délicate appréhension de l'incertitude lorsque la population est observée exhaustivement ;
- Rôle de la data vizualisation dans la communication auprès du grand public.

Conclusions

- L'INSEE communique autour de l'incertitude des statistiques publiques, et une évolution historique vers plus de transparence est notable ;
- Communication limitée à des indicateurs phares dans les publications à diffusion large et pour la vulgarisation scientifique des méthodes, mais beaucoup plus précise et détaillée dans des publications académiques ou spécialisées ;
- L'effort fourni à partir de 2008 pour mieux communiquer les incertitudes entourant les prévisions de croissance visait aussi à alerter le public sur la distinction entre prévision et statistique ;
- La crise de reproductibilité des sciences sociales pourrait déboucher sur une meilleure information sur l'incertitude entourant les données détaillées accessibles aux chercheurs.

Conclusions

- L'incertitude recouvre plusieurs formes, quantifiables ou non, permanente, transitoire ou conceptuelle ;
- Communiquer sur l'incertitude, c'est faire un choix, qui peut lui-même donner l'illusion d'exactitude ;
- Il interfère aussi avec le jugement de la crédibilité des statistiques, indispensable pour leur acceptation dans le débat public ;
- L'arbitrage entre transparence, pédagogie et défence de l'institution demeure délicat ;
- A l'heure de la "data visualization", de nouveaux modes de communication de l'incertitude pourraient être réfléchis.