**Licence AES 3ème année**

**Université de Lille II**

**Notes du cours d'introduction à l'économétrie**

**P. De Vreyer**

## Partie II: Notions d'inférence statistique.

Supposons que l'on désire connaître la moyenne des revenus dans une population donnée, par exemple la population française. Comment peut-on s'y prendre ? Le plus simple serait d'interroger tous les membres de la population, de demander à chacun son revenu et d'en faire la moyenne. Mais cela coûterait trop cher. En général on doit se contenter d'interroger une fraction de la population (ce que l'on appelle un échantillon) et de se contenter de l'information recueillie auprès de ces personnes. Le problème auquel fait face le statisticien est le suivant:

> *Comment, à partir de l'information recueillie auprès d'un échantillon de la population, en inférer des conclusions qui valent pour l'ensemble de la population ?*

Dans notre exemple, la question est d'autant plus difficile que la véritable valeur du revenu est, et restera, inconnue du statisticien.

Pour y parvenir il faut établir une « stratégie », c'est à dire décider de la façon dont on va « deviner » la véritable valeur du revenu moyen, à partir des informations contenues dans l'échantillon. Nous verrons qu'il y a, *a priori*, une infinité de « stratégies » possibles, mais que seul un faible nombre d'entre-elles sont bonnes. L'objet de cette partie est de vous présenter les critères de choix entre les stratégies possibles. Auparavant un peu de vocaculaire doit être introduit.

### 1. Quelques définitions

Supposons que l'on désire évaluer un ensemble de paramètres (moyenne, variance etc.) de la loi d'un vecteur aléatoire X.

> Définition 1: Observation.
> On appelle observation une valeur du vecteur X « observée » par l'économètre.
>
> Définition 2: Population.
> La population est constituée de l'ensemble des observations qui peuvent être faites par l'économètre.
>
> Définition 3: Echantillon indépendant.
> Un échantillon indépendant est un sous ensemble de la population obtenu par tirage aléatoire des observations dans lequel on suppose que toutes les observations sont indépendantes les unes des autres.
>
> Dans toute la suite de ce cours on supposera que tous les échantillons dont il est question sont indépendants.
>
> Exemple: Chaque année l'INSEE réalise au mois de Mars une enquête nationale sur l'emploi. La population concernée est ici l'ensemble de la population active française. Les ménages qui sont interrogés dans cette enquête constituent un échantillon d'observations d'une taille égale à environ 80000 ménages. Pour chaque ménage un ensemble d'informations sont

enregistrées qui peuvent être rangées dans un vecteur. L'ensemble des vecteurs qui en résulte constitue l'échantillon des observations dont dispose l'économètre. Cet échantillon est indépendant car les informations recueillies auprès d'un ménage quelconque ne dépendent pas de celles recueillies auprès d'un autre ménage. Par exemple, le fait que monsieur Dupont à Marseille soit ingénieur est indépendant du fait que monsieur Durand à Brest soit vendeur de chaussures.

*Pour la suite, il est important de bien comprendre que, du point de vue de l'économètre la valeur du vecteur des observations de monsieur Dupont est aléatoire, même si elle ne l'est pas pour monsieur Dupont lui-même.* Quand l'économètre descend dans la rue pour interroger monsieur Dupont et lui poser des questions sur son emploi, il ne sait pas *a priori* ce que va lui répondre monsieur Dupont. Les réponses de monsieur Dupont sont donc, du point de vue de l'économètre, des observations d'un vecteur aléatoire.

Définition 4: Estimateur.
Soit $X_{1,}\ldots, X_n$ un n-uplet de variables aléatoires indépendantes suivant toutes la loi d'une variable aléatoire X. Un estimateur est une fonction de $X_{1,}\ldots, X_n$.

Définition 5: Estimation.
La valeur prise par un estimateur sur un échantillon particulier est appelée estimation.

Définition 6: Moyenne empirique.
Soit $X_{1,}\ldots, X_n$ un n-uplet de variables aléatoires indépendantes suivant toutes la loi d'une variable aléatoire X. La moyenne empirique est un estimateur dont l'expression est donnée par:

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n} X_i$$

La moyenne empirique peut-être utilisée pour estimer la moyenne de la loi de X. Soit $x=(x_{1,}..,x_n)'$ un échantillon d'observations de X. L'estimation de la moyenne de X par la moyenne empirique est donnée par:

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

Exemple: on observe le revenu X d'un échantillon de 80000 ménages tiré de la population française (qui en comporte plusieurs millions). Soit $X_i$ la variable aléatoire égale au revenu du ménage i. Du point de vue de l'économètre, qui ne connaît pas *a priori* quelle peut-être la valeur de X, $X_i$ est une variable aléatoire. La stratégie que peut employer l'économètre pour estimer la véritable valeur du revenu moyen de la population est de calculer l'estimateur de la moyenne empirique. La valeur qu'il obtient pour l'échantillon dont il dispose est une estimation de la vraie moyenne.

Définition 7: Variance empirique.
Soit $X_{1,}\ldots, X_n$ un n-uplet de variables aléatoires indépendantes suivant toutes la loi d'une variable aléatoire X. La variance empirique est un estimateur dont l'expression est donnée par:

$$\sigma_X{}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$
$$= \frac{1}{n}\sum_{i=1}^{n} X_i{}^2 - \bar{X}^2$$

Pour un échantillon donné la valeur de cet estimateur est donnée par:

$$\sigma_x{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
$$= \frac{1}{n}\sum_{i=1}^{n} x_i{}^2 - \bar{x}^2$$

Définition 8: Covariance empirique.
Soit $(X_1, Y_1), \ldots, (X_n, Y_n)$ un n-uplet de couples de variables aléatoires indépendantes suivant toutes la loi du couple $(X, Y)$. La covariance empirique est un estimateur dont l'expression est donnée par:

$$\sigma_{XY} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - \bar{X})(Y_j - \bar{Y})$$
$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i Y_j - \bar{X}\,\bar{Y}$$

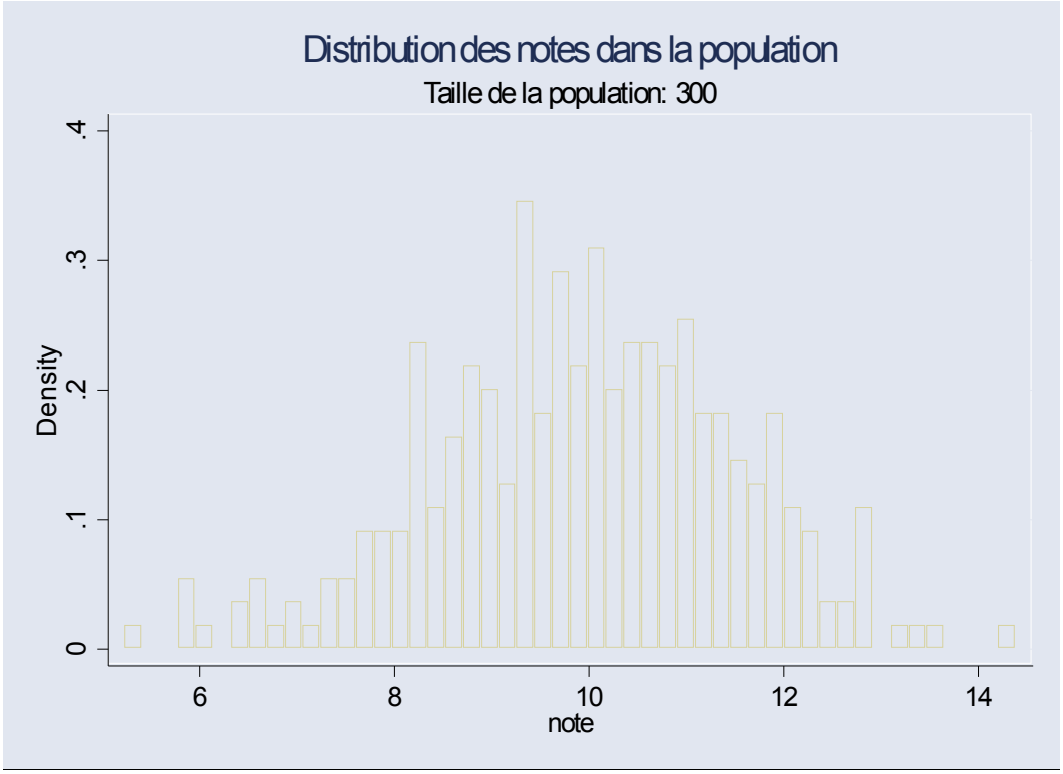
Pour un échantillon donné la valeur de cet estimateur est donnée par:

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \bar{x})(y_j - \bar{y})$$
$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j - \bar{x}\,\bar{y}$$

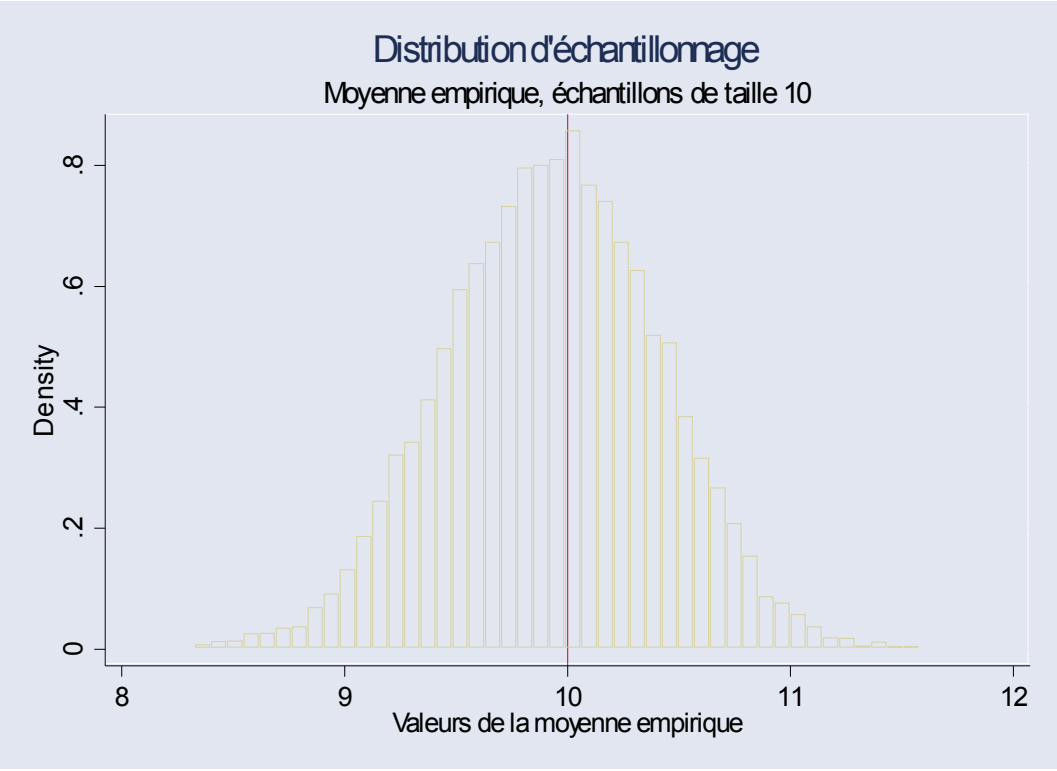
Définition 9: Distribution d'échantillonnage.
La valeur d'un estimateur peut être calculée pour chaque échantillon que l'on peut tirer dans la population. La distribution d'échantillonnage d'un estimateur est la distribution de l'ensemble des valeurs possibles de cet estimateur, calculées sur l'ensemble des échantillons qui peuvent être tirés dans cette population.

Exemple: Supposons que la population soit celle des étudiants en première année AES. Il y a 300 étudiants dans cette population. La variable aléatoire est la note obtenue au partiel de macro-économie. Pour estimer la moyenne on tire un échantillon de 10 copies d'étudiants parmi les 300. Le nombre d'échantillons possibles est donné par: $C(300,10) = \frac{300!}{10!(290!)} = 1{,}39832\ 10^{18}$. Pour chaque échantillon, on pourrait imaginer de calculer la moyenne empirique des notes obtenues par les 10 étudiants tirés. Cependant, cela ne signifie pas que le nombre de valeurs possibles prises par la moyenne empirique soit aussi important que C(300,10), parce que plusieurs étudiants peuvent avoir la même note. La distribution d'échantillonnage est la distribution des valeurs de la moyenne empirique que l'on obtiendrait si l'on tirait chacun des C(300,10) échantillons.

La figure ci-dessous montre la distribution des notes de la population des 300 étudiants:




Et voici maintenant la distribution d'échantillonnage:




J'ai supposé dans cet exemple que la vraie moyenne des notes obtenues à l'examen est égale à 10. Vous pouvez remarquer que les deux distributions sont centrées autour de cette valeur. Ceci est particulièrement remarquable pour la distribution d'échantillonnage. Il s'agit là d'un des deux critères généralement retenu pour choisir les estimateurs que l'on calcule sur un échantillon.

## 2. Critères de choix des estimateurs.

Définition 10: Absence de biais.

L'absence de biais est un des critères que l'on retient en général pour choisir les estimateurs. Soit $X_1, \ldots, X_n$ un n-uplet de variables aléatoires distribuées selon la loi de X. Soit $\mu$ un paramètre de cette loi (par exemple sa moyenne). Soit $g(X_1, .., X_n)$ un estimateur de $\mu$. On dit que g est sans biais si: $E(g(X_1, .., X_n)) = \mu$.

Exemple: Si g est la moyenne empirique et $\mu$ est la moyenne de X, alors g est un estimateur sans biais de $\mu$. En effet:

$$E(g(X_1, .., X_n)) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

puisque chacune des variables aléatoires $X_i$ est distribuée selon la loi de X.

Pourquoi l'absence de biais est-elle un critère qu'il est souhaitable de vérifier lorsque l'on choisit un estimateur ? Il faut bien comprendre la situation de l'économètre: *lorsqu'il doit estimer un paramètre un seul échantillon est à sa disposition*. Par conséquent, pour estimer la valeur du paramètre recherché il va choisir une stratégie, autrement dit un estimateur, qui « marche » en moyenne, autrement dit un estimateur sans biais.

Définition 11: Efficacité.

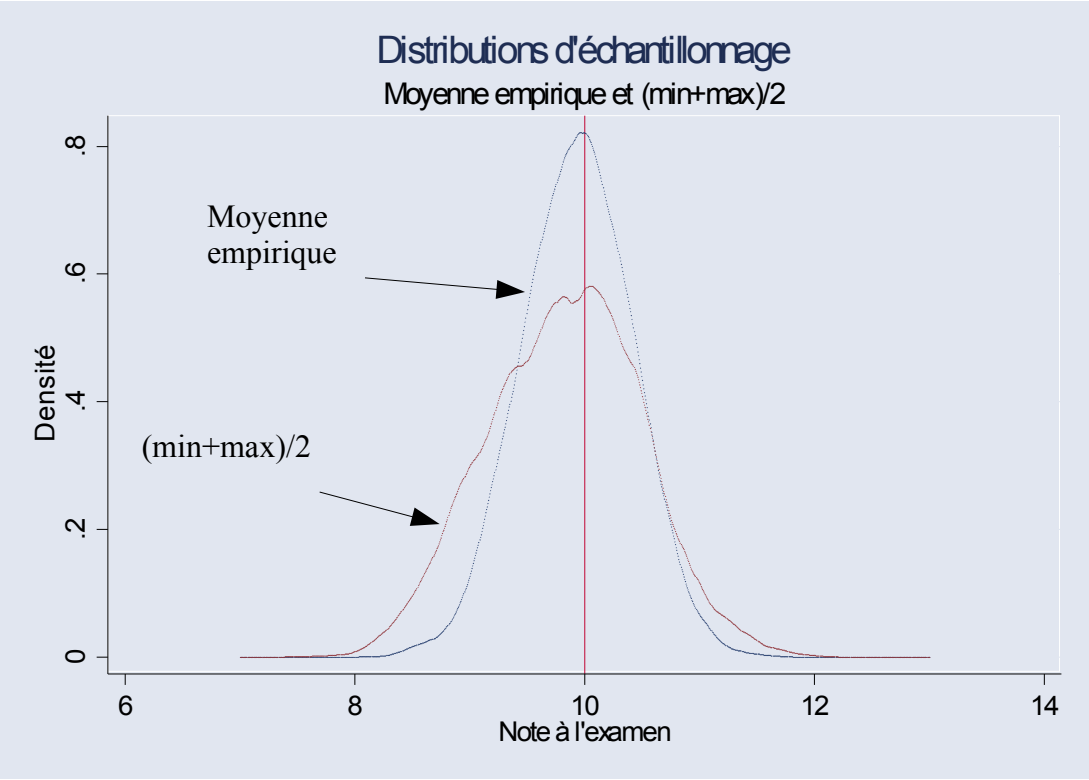
C'est le second critère que l'on retient en général. Soit $X_1, \ldots, X_n$ un n-uplet de variables aléatoires distribuées selon la loi de X. Soit $\mu$ un paramètre de cette loi (par exemple sa moyenne). Soient $g_1(X_1, .., X_n)$ et $g_2(X_1, .., X_n)$ deux estimateurs de $\mu$. On dit que $g_1$ est plus efficace que $g_2$ lorsque $V(g_1) < V(g_2)$.

L'estimateur qui a la variance la plus faible est appelé l'estimateur efficace.

L'efficacité est un critère que l'on désire vérifier, encore une fois parce que l'économètre ne dispose que d'un seul échantillon. Il faut donc qu'il puisse faire confiance à l'estimation obtenue avec cet échantillon. Pour cela il choisit la stratégie d'estimation qui non seulement « marche » en moyenne, mais qui en plus minimise le risque que, sur un échantillon particulier, la valeur obtenue soit très éloignée de la vraie valeur du paramètre. C'est là le sens du critère d'efficacité.

Exemple: Le graphique ci-dessous montre la distribution d'échantillonnage de deux estimateurs non biaisés de la moyenne. Le premier est la moyenne empirique calculée sur des échantillons de 10 étudiants. Le second est la moyenne de la note maximale et de la note minimale observée sur ces mêmes échantillons. Les deux estimateurs sont sans biais. Cependant la moyenne empirique est un estimateur plus efficace: sa distribution est plus resserrée autour de la moyenne et les valeurs éloignées ont une probabilité plus faible d'être loin de la vraie valeur.

Il est facile de montrer ces résultats:




$E[\frac{(min+max)}{2}]=\frac{1}{2}\left(E[min(X_1,\ldots,X_n)]+E[max(X_1,\ldots,X_n)]\right)=\mu$ puisque toutes les variables aléatoires $X_i$ ont la même distribution de moyenne $\mu$.
Pour les variances:

$V(\bar{X})=V\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}V(X_i)=\frac{\sigma_X{}^2}{\mathrm{n}}$ puisque les variables aléatoires $X_i$ sont indépendantes.

$V\left(\frac{(min+max)}{2}\right)=\frac{1}{4}\left(V(min(X_1,\ldots,X_n))+V(max(X_1,\ldots,X_n))\right)=\frac{\sigma_X{}^2}{2}$

pour la même raison que précédemment.

On voit donc que parmi les deux estimateurs, la moyenne empirique est le plus efficace.