Introduction à l'économétrie

WWW.ECONOMIE-GESTION.COM

Bref aperçu du cours

La théorie économique suggère des relations entre variables, qui ont souvent des implications politiques, mais ne dit pratiquement jamais rien sur l'importance quantitative de ces effets causaux

- Quel est l'effet *quantitatif* de la réduction de la taille des classes sur la réussite des élèves?
- Quel est l'effet sur le salaire d'une année supplémentaire d'éducation?
- Quelle est l'élasticité-prix de la cigarette?
- Quel est l'effet sur la croissance du PNB d'une hausse de 1 point des taux d'intérêt de la banque centrale ?
- Quel est l'effet sur les prix du logement de l'amélioration de l'environnement?

WWW.ECONOMIE-GESTION.COM

Dans ce cours on utilisera des données pour mesurer des effets causaux.

- Idéalement, on aimerait pouvoir se baser sur une expérience
  - Quelle serait l'expérience qui permet d'estimer l'effet d'une réduction de la taille des classes sur des notes d'examen des étudiants?
- Malheureusement, on a presque toujours accès uniquement à des données observationnelles (et non expérimentales).
  - rendements de l'éducation
  - prix des cigarettes
  - politique monétaire
- Dans ce cours nous verrons les difficultés que pose l'estimation d'effets causaux à partir de données observationnelles
  - effets confondants (facteurs omis)
  - causalité simultanée
  - "corrélation n'est pas causalité"

WWW.ECONOMIE-GESTION.COM

Dans ce cours:

- On étudiera les méthodes d'estimation des effets causaux à partir de données d'observation
- On étudiera des outils qui peuvent être utilisés à d'autres fins, par exemple la prévision à partir des données de séries chronologiques;
- On mettra l'accent sur les applications –on ne fait appel à la théorie que lorsqu'elle est nécessaire à la compréhension du pourquoi des méthodes;
- On apprendra à évaluer les analyses de régression des autres - ce qui signifie que vous serez capables de lire et comprendre des travaux empiriques en économie dans d'autres cours d'économie;
- Vous apprendrez la pratique de la régression dans vos séries d'exercices.

WWW.ECONOMIE-GESTION.COM

Rappels de Probabilité et Statistique

Problème empirique: Taille de la classe et réussite scolaire

- Question de politique: Quel est l'effet sur les résultats scolaires (ou sur une autre mesure) d'une réduction de la taille des classes d'un étudiant par classe? de 8 étudiant s/classe?
- Il faut avoir recours à des données pour avoir la réponse (peut-on répondre à cette question *sans* données?)

WWW.ECONOMIE-GESTION.COM

Les données de tests standards de Californie

Tous les districts scolaires de Californie (n = 420)

Variables:

- Résultats de tests standards de 5e année (Stanford-9 achievement test : test combiné de math et de lecture), moyenne par district
- Rapport élève/enseignant (STR) = no. d'élèves dans le district divisé par le no d'équivalents temps-plein d'enseignants

WWW.ECONOMIE-GESTION.COM

Un premier coup d'œil aux données:

(*Vous devriez être capables d'interpréter cette table*)

TABLE 4.1 Summary of the Distribution of Student–Teacher Ratios and Fifth-Grade Test Scores for 420 K–8 Districts in California in 1998

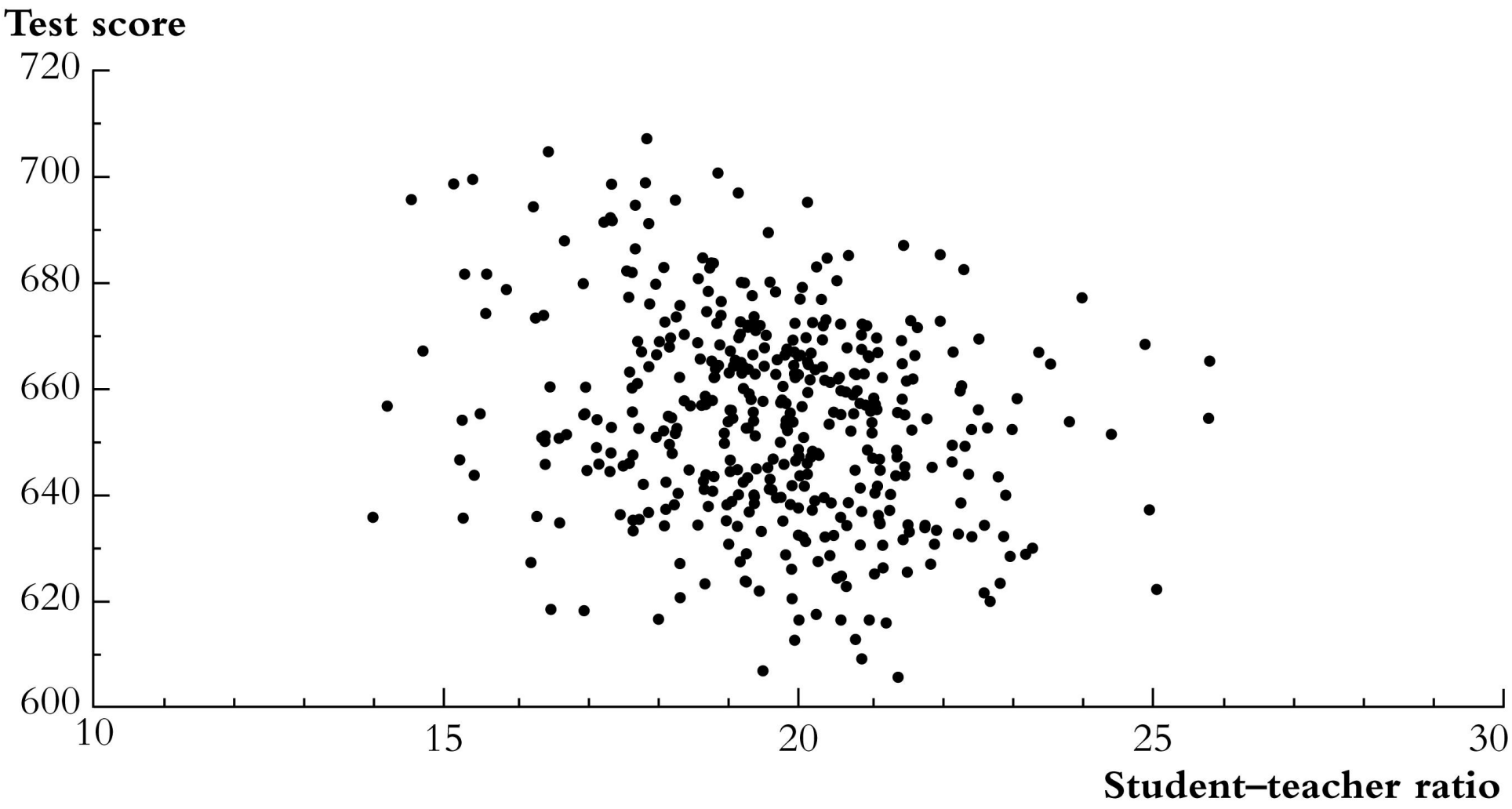
			Percentile						
	Average	Standard Deviation	10%	25%	40%	50% (median)	60%	75%	90%
Student–teacher ratio	19.6	1.9	17.3	18.6	19.3	19.7	20.1	20.9	21.9
Test score	665.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

Cette table ne nous dit rien sur la relation entre les résultats de tests et la variable *STR*.

WWW.ECONOMIE-GESTION.COM

Est-ce que les districts avec des classes plus petites ont de meilleurs résultats de test?

Scatterplot des résultats aux test vs. élèves/enseignant



Que montre ce graphique?

WWW.ECONOMIE-GESTION.COM

Nous cherchons à établir des faits sur la base d'informations numériques pour savoir si les districts avec un STR bas ont de meilleurs résultats de test – mais comment ?

1. Comparer les notes moyennes de test dans les districts avec un STR bas à ceux des districts avec un STR élevé ("***estimation***")

2. Tester l'hypothèse "nulle" que la moyenne des résultats des tests dans les deux types de districts sont les mêmes contre l'hypothèse "alternative" qu'ils sont différents ("***test d'hypothèse***")

3. Estimer un intervalle pour la différence dans les résultats moyens des tests pour les districts avec STR élevés et bas ("***intervalle de confiance***")

WWW.ECONOMIE-GESTION.COM

Analyse de données initiale: Comparons les districts avec classes "petites" (STR < 20) et "grandes" (STR ≥ 20):

Taille Classes	Résultat moyen ($\bar{Y}$)	Ecart-type (s_Y)	n
Petite	657.4	19.4	238
Grande	650.0	17.9	182

1. ***Estimation*** de Δ = différence entre les moyennes de groupes
2. ***Test de l'hypothèse*** que $\Delta = 0$
3. Construction d'un ***intervalle de confiance*** pour Δ

WWW.ECONOMIE-GESTION.COM

1. Estimation

$$\bar{Y}_{\text{small}} - \bar{Y}_{\text{large}} = \frac{1}{n_{\text{small}}} \sum_{i=1}^{n_{\text{small}}} Y_i - \frac{1}{n_{\text{large}}} \sum_{i=1}^{n_{\text{large}}} Y_i$$

$$= 657.4 - 650.0$$

$$= 7.4$$

Est-ce que du point de vue pratique c'est une différence importante?

- Ecart-type entre les districts = 19.1
- Différence entre le 60-ième et le 75-ième percentiles de la distribution des résultats de test est 667.6 – 659.4 = 8.2
- Est-ce suffisamment grand pour être important dans une discussion sur une réforme scolaire?

WWW.ECONOMIE-GESTION.COM

2. Test d'hypothèse

Test de la différence de moyennes: calculez la *t*-stat,

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{\bar{Y}_s - \bar{Y}_l}{SE(\bar{Y}_s - \bar{Y}_l)}$$ (vous vous souvenez?)

avec $SE(\bar{Y}_s - \bar{Y}_l)$ l'écart-type" de $\bar{Y}_s - \bar{Y}_l$, les indices *s* et *l* pour les districts avec STR "petits (small)" et "large (grand)", et

$$s_s^2 = \frac{1}{n_s - 1}\sum_{i=1}^{n_s}(Y_i - \bar{Y}_s)^2 \text{ (etc.)}$$

WWW.ECONOMIE-GESTION.COM

Calculez la *t*-stat pour la différence de moyennes:

Taille	$\bar{Y}$	s_Y	n
petite	657.4	19.4	238
grande	650.0	17.9	182

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{657.4 - 650.0}{\sqrt{\frac{19.4^2}{238} + \frac{17.9^2}{182}}} = \frac{7.4}{1.83} = 4.05$$

|t| > 1.96, donc on rejette l'hypothèse nulle que les deux groupes ont la même moyenne (à un seuil de 5%)

WWW.ECONOMIE-GESTION.COM

3. Intervalle de confiance

Un intervalle de confiance à 95% pour la différence entre les moyennes est,

$$(\bar{Y}_s - \bar{Y}_l) \pm 1.96 \times SE(\bar{Y}_s - \bar{Y}_l)$$

$$= 7.4 \pm 1.96 \times 1.83 = (3.8, 11.0)$$

Deux affirmations équivalentes:

1. L'intervalle de confiance à 95% pour Δ ne contient pas 0;
2. L'hypothèse que $\Delta = 0$ est rejetée au seuil de 5%.

WWW.ECONOMIE-GESTION.COM

La suite…

- L'aspect mécanique de l'estimation, des tests d'hypothèse, et des intervalles de confiances devrait être connu
- Ces concepts peuvent être étendus tels quels à la régression et à ses variations
- Avant de passer à la régression, nous allons revoir un peu la théorie sur laquelle sont fondés l'estimation, les tests d'hypothèse, et les intervalles de confiance:
- Pourquoi ces méthodes fonctionnent-elles, et pourquoi utiliser une méthode plutôt qu'une autre ? Nous allons revoir les fondements de la statistique et de l'économétrie

WWW.ECONOMIE-GESTION.COM

Rappels de Statistique

1. **Le cadre probabiliste de l'inférence statistique**
2. Estimation
3. Tests d'hypothèse
4. Intervalles de Confiance

<u>Cadre probabiliste pour l'inférence statistique</u>

(a) Population, variable aléatoire et distribution

(b) Moments d'une distribution (moyenne, variance, écart-type, covariance, corrélation)

(c) Distributions conditionnelles et moyenne conditionnelle

(d) Distribution d'un échantillon de données tirées au hasard d'une population: $Y_1, \ldots, Y_n$

WWW.ECONOMIE-GESTION.COM

(a) Population, variable aléatoire et distribution

Population

- Le groupe ou l'ensemble de toutes les unités qui nous intéressent (districts scolaires)
- Nous ferons comme si la population avait une taille infinie (∞ est une approximation pour "très grand")

Variable aléatoire Y

- Résumé numérique d'un événement aléatoire (résultats moyens d'examen dans un district)

WWW.ECONOMIE-GESTION.COM

Distribution de Y dans la Population

- Les probabilités de chacune des différentes valeurs que peut prendre Y dans la population, par ex. $P[Y = 650]$ (quand Y est discrète)
- ou: Les probabilités d'ensembles de valeurs, par ex. $P[640 \leq Y \leq 660]$ (quand Y est continue).

WWW.ECONOMIE-GESTION.COM

(b) Moments de la distribution d'une population: moyenne, variance, écart-type, covariance, corrélation

moyenne = valeur espérée (espérance) de Y

$= E(Y)$

$= \mu_Y$

= moyenne à long terme de Y sur un grand nombre de réalisations répétées de Y

variance $= E(Y - \mu_Y)^2$

$= \sigma_Y^2$

= mesure les déviations de la distribution par rapport à la moyenne, au carré

écart-type $= \sqrt{\text{variance}} = \sigma_Y$

WWW.ECONOMIE-GESTION.COM

Moments, suite.

$$\textbf{\textit{skewness}} = \frac{E\left[\left(Y-\mu_Y\right)^3\right]}{\sigma_Y^3}$$

= mesure l'asymétrie de la distribution

- *skewness* = 0: distribution symétrique
- *skewness* > (<) 0: la distribution a une longue queue à droite (à gauche)

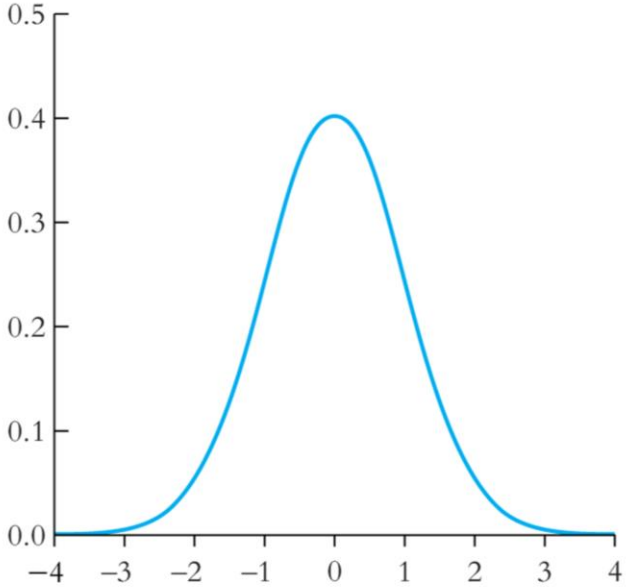
$$\textbf{\textit{kurtosis}} = \frac{E\left[\left(Y-\mu_Y\right)^4\right]}{\sigma_Y^4}$$

= mesure masse dans les queues de distribution

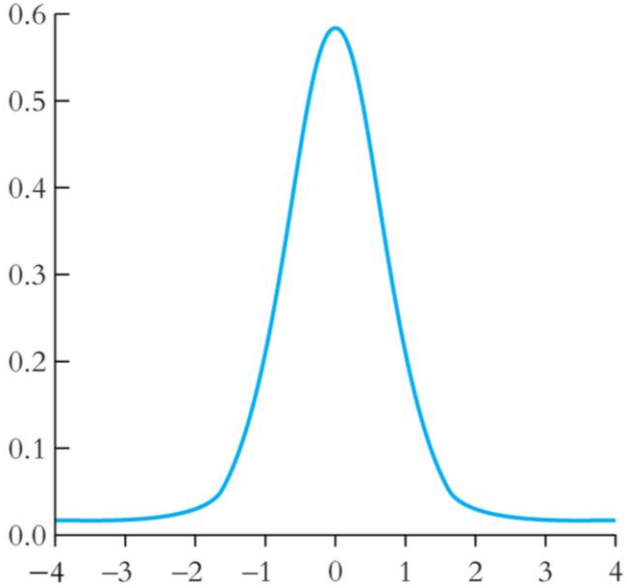
= mesure la probabilité de grandes valeurs

- *kurtosis* = 3: distribution normale
- *skewness* > 3: queues épaisses ("***leptokurtose***")

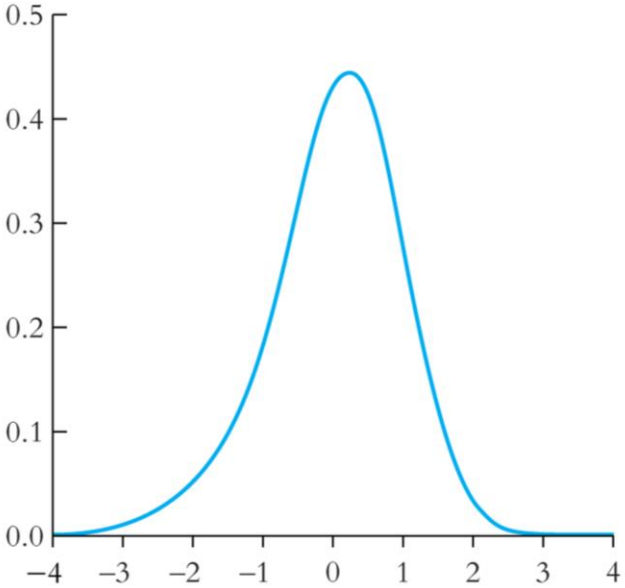
WWW.ECONOMIE-GESTION.COM



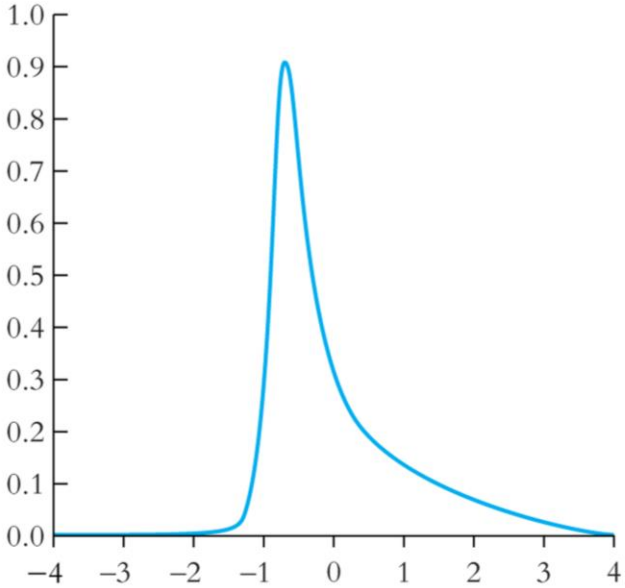
(a) Skewness = 0, kurtosis = 3



(b) Skewness = 0, kurtosis = 20



(c) Skewness = −0.1, kurtosis = 5



(d) Skewness = 0.6, kurtosis = 5

WWW.ECONOMIE-GESTION.COM

2 variables aléatoires: distributions jointes et covariance

- Les variables **aléatoires** X et Z ont une ***distribution jointe***
- La ***covariance*** entre X et Z est

$$\text{cov}(X,Z) = E[(X - \mu_X)(Z - \mu_Z)] = \sigma_{XZ}$$

- La covariance est une mesure linéaire de l'association entre X et Z; ses unités sont les unités de $X \times$ unités de Z
- $\text{cov}(X,Z) > 0$ il existe une relation positive entre X et Z
- Si X et Z sont distribuées indépendamment, alors $\text{cov}(X,Z) = 0$ (mais pas l'inverse!!)
- La covariance d'une v.a. avec elle-même est la variance:

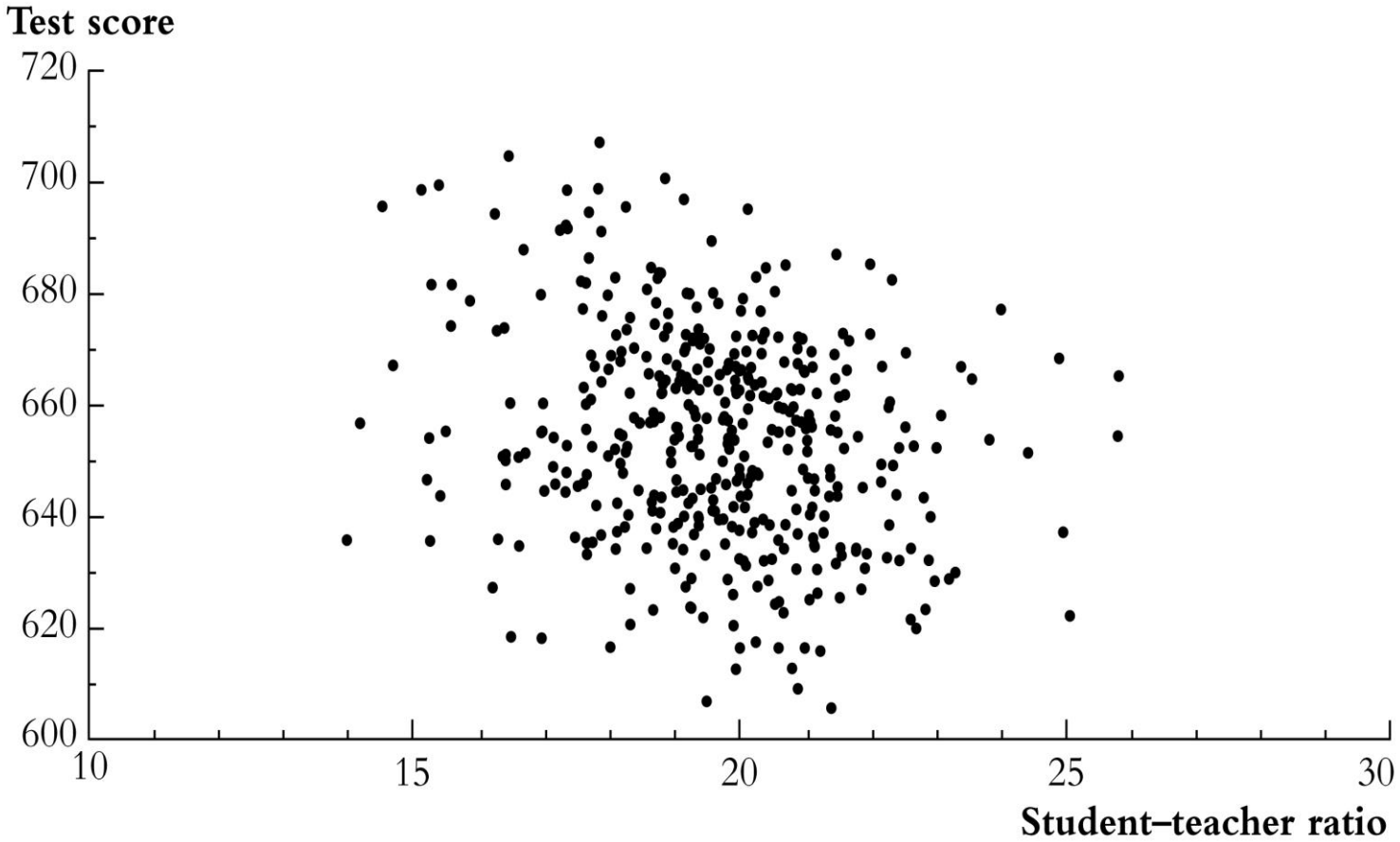
$$\text{cov}(X,X) = E[(X - \mu_X)(X - \mu_X)] = E[(X - \mu_X)^2] = \sigma_X^2$$

WWW.ECONOMIE-GESTION.COM

La covariance entre *Test Score* et *STR* est négative:

FIGURE 4.2 **Scatterplot of Test Score vs. Student–Teacher Ratio (California School District Data)**

Data from 420 California school districts. There is a weak negative relationship between the student–teacher ratio and test scores: The sample correlation is −0.23.



La ***corrélation aussi***…

WWW.ECONOMIE-GESTION.COM

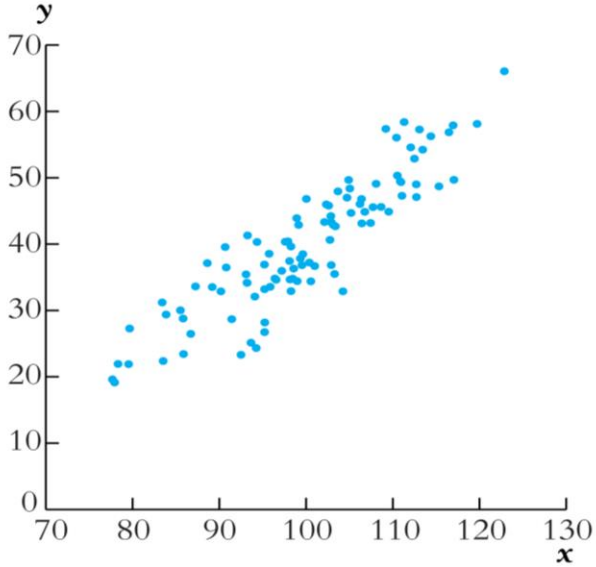
Le ***coefficient de corrélation*** est défini en fonction de la covariance:

$$\text{corr}(X,Z) = \frac{\text{cov}(X,Z)}{\sqrt{\text{var}(X)\,\text{var}(Z)}} = \frac{\sigma_{XZ}}{\sigma_X \sigma_Z} = r_{XB}$$

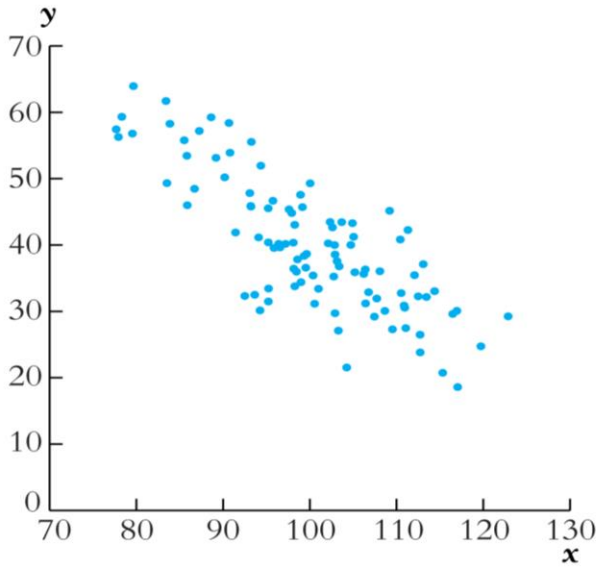
- $-1 \leq \text{corr}(X,Z) \leq 1$
- $\text{corr}(X,Z) = 1$ association linéaire positive parfaite
- $\text{corr}(X,Z) = -1$ association linéaire négative parfaite
- $\text{corr}(X,Z) = 0$ pas d'association linéaire

WWW.ECONOMIE-GESTION.COM

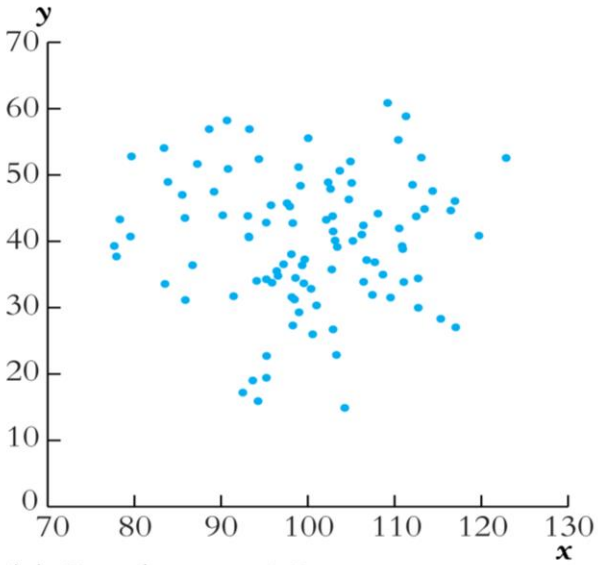
Le coefficient de corrélation mesure l'association linéaire



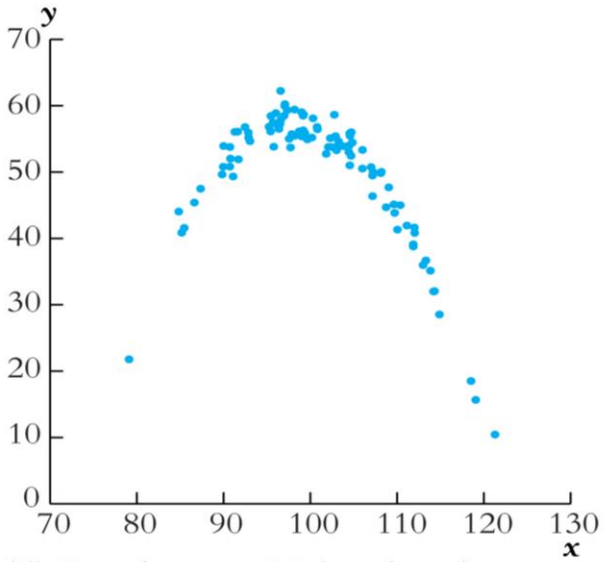
(a) Correlation = +0.9



(b) Correlation = –0.8



(c) Correlation = 0.0



(d) Correlation = 0.0 (quadratic)

WWW.ECONOMIE-GESTION.COM

(c) Distributions et moyennes conditionnelles

Distributions conditionnelles

- La distribution de Y, étant donné la (ou les) valeur(s) d'une autre variable aléatoire, X
- Ex: la distribution des résultats de test, étant donné que STR < 20

Espérance et Moments Conditionnels

moyenne conditionnelle = moyenne de la distribution conditionnelle

$$= E(Y|X = x)$$ (***concept et notation importants!***)

- *variance conditionnelle* = variance de la distribution conditionnelle

WWW.ECONOMIE-GESTION.COM

- *Exemple*: $E(Test\ scores|STR < 20)$ = moyenne des résultats de test dans les districts avec de petites classes

La différence de moyennes est la différence entre les moyennes des deux distributions conditionnelles:

Moyenne conditionnelle, suite.

$\Delta = E(Test\ scores|STR < 20) - E(Test\ scores|STR \geq 20)$

Autres exemples de moyennes conditionnelles:

- Les salaires des femmes (Y = salaires, X = sexe)
- Le taux de mortalité de ceux qui reçoivent un traitement expérimental (Y = vivre/mourir; X = traité/pas traité)
- Si $E(X|Z)$ = const, alors corr(X,Z) = 0 (mais pas forcément vice versa)

La moyenne conditionnelle est un terme technique qui recouvre l'idée familière de moyenne par groupe

WWW.ECONOMIE-GESTION.COM

(d) Distribution d'un échantillon de données pris au hasard d'une population: $Y_1,\ldots, Y_n$

Nous ferons l'hypothèse d'un tirage aléatoire simple

- Choisissons un individu (district, entité) au hasard dans la population

Aléa et données

- Avant que l'échantillon ne soit tiré, la valeur de Y est aléatoire, parce que l'individu est choisi au hasard
- Une fois que l'individu a été choisi, la valeur de Y est observée, et alors Y est juste un nombre – il n'est plus aléatoire
- Les données sont (Y_1, $Y_2,\ldots, Y_n$), où Y_i = valeur de Y pour le i-ième individu (district, unité) échantillonné

WWW.ECONOMIE-GESTION.COM

Distribution de $Y_1,\ldots, Y_n$ avec échantillonnage aléatoire simple

- Comme les individus 1 et 2 sont choisis au hasard, la valeur de Y_1 ne donne aucune information concernant Y_2. Donc:
  - Y_1 and Y_2 sont ***indépendamment distribués***
  - Y_1 and Y_2 viennent de la même distribution, càd que Y_1, Y_2 sont ***identiquement distribués***
  - Donc, avec un échantillonnage aléatoire simple, Y_1 et Y_2 sont indépendamment et identiquement distribués (***i.i.d.***).
  - De façon générale, avec un échantillonnage aléatoire simple, $\{Y_i\}$, $i = 1,\ldots, n$, sont i.i.d.

Ce cadre permet de réaliser une inférence statistique rigoureuse sur les moments de la distribution de la population à partir d'un échantillon de données pris de cette population…

WWW.ECONOMIE-GESTION.COM

1. Le cadre probabiliste de l'inférence statistique
2. **Estimation**
3. Tests d'hypothèse
4. Intervalles de Confiance

<u>Estimation</u>

$\bar{Y}$ est l'estimateur naturel de la moyenne. Mais:

(a) Quelles sont les propriétés de $\bar{Y}$?

(b) Pourquoi utiliser $\bar{Y}$ plutôt qu'un autre estimateur?

- Y_1 (la première observation)?
- Peut-être des pondérations différentes – pas juste la moyenne simple ?
- La médiane(Y_1,…, Y_n)?

Le point de départ est la distribution d'échantillonnage de $\bar{Y}$ …

WWW.ECONOMIE-GESTION.COM

(a) Distribution d'échantillonnage de $\bar{Y}$

$\bar{Y}$ est une variable aléatoire, et ses propriétés sont déterminées par la **distribution d'échantillonnage de** $\bar{Y}$

- Les individus dans l'échantillon sont tirés au hasard.
- Donc les valeurs de $(Y_1,\ldots, Y_n)$ sont aléatoires
- Donc des fonctions de $(Y_1,\ldots, Y_n)$, telles que $\bar{Y}$, sont aléatoires: si un échantillon différent avait été tiré, ils prendraient des valeurs différentes
- La distribution de $\bar{Y}$ sur les différents échantillons possibles de taille n s'appelle la **distribution d'échantillonnage** de $\bar{Y}$.
- La moyenne et la variance de $\bar{Y}$ sont les moyenne et variance de sa distribution d'échantillonnage, $E(\bar{Y})$ et $\text{var}(\bar{Y})$.
- Le concept de distribution d'échantillonnage est à la base de toute l'économétrie.

WWW.ECONOMIE-GESTION.COM

La distribution d'échantillonnage de $\bar{Y}$***, suite.***

Exemple: Supposons que Y prend comme valeurs 0 ou 1 (une variable aléatoire ***Bernoulli***) avec distribution,

$$\Pr[Y = 0] = .22,\ \Pr(Y = 1) = .78$$

Alors

$$E(Y) = p\times 1 + (1 - p)\times 0 = p = .78$$

$$\sigma_Y^2 = E[Y - E(Y)]^2 = p(1 - p) \quad \textit{[Vous vous souvenez?]}$$

$$= .78\times(1-.78) = 0.1716$$

La distribution d'échantillonnage de $\bar{Y}$ dépend de n.

Supposons $n = 2$. La distribution d'échantillonnage de $\bar{Y}$ est,

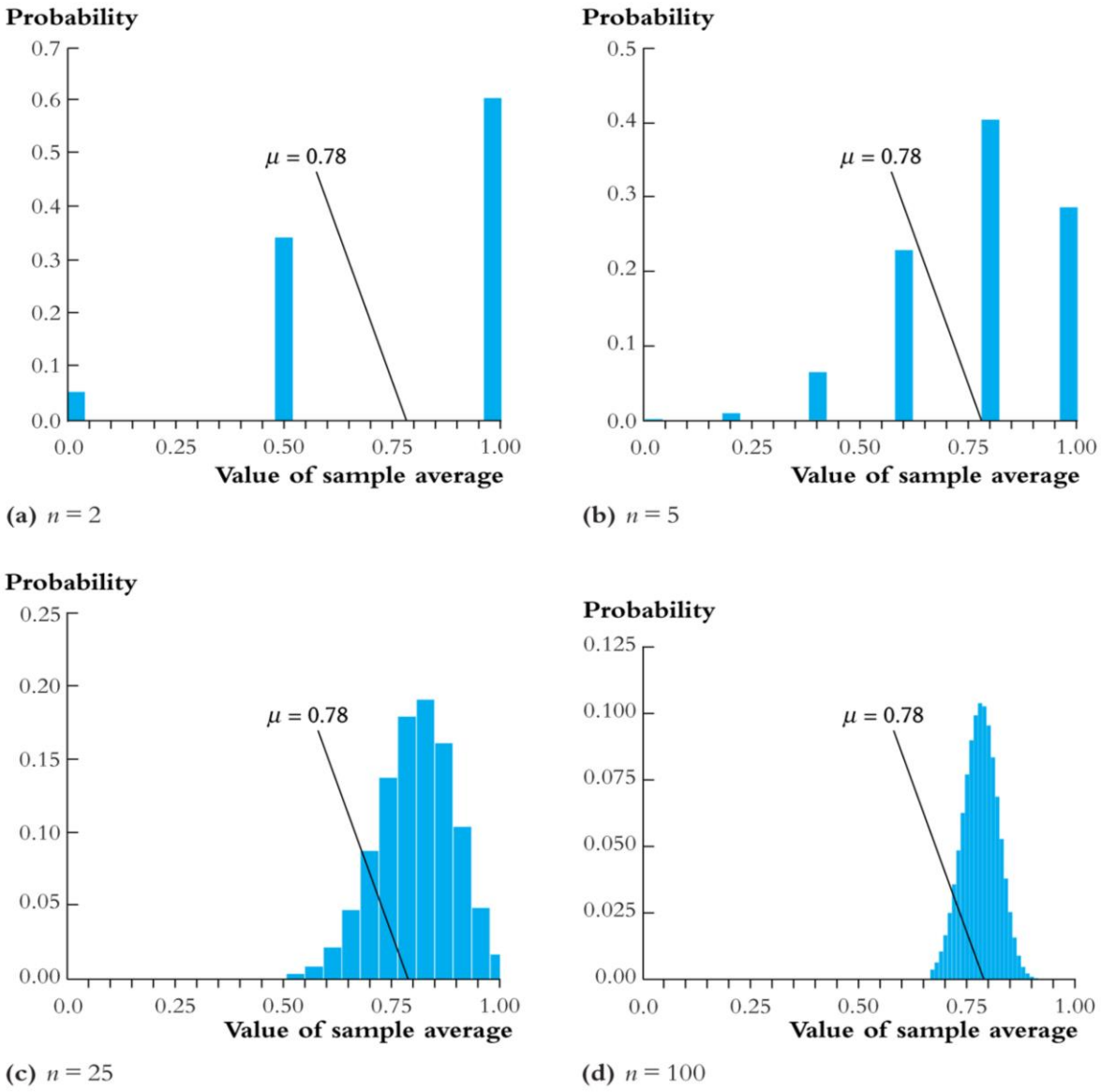
$$\Pr(\bar{Y} = 0) = .22^2 = .0484$$

$$\Pr(\bar{Y} = ½) = 2\times.22\times.78 = .3432$$

$$\Pr(\bar{Y} = 1) = .78^2 = .6084$$

WWW.ECONOMIE-GESTION.COM

Distribution d'échantillonnage de $\bar{Y}$ avec Y Bernoulli (p=0.78)



WWW.ECONOMIE-GESTION.COM

Que doit-on savoir de la distribution d'échantillonnage:

- Quelle est la moyenne de $\bar{Y}$?
  - Si $E(\bar{Y})$ = vrai μ = .78, alors $\bar{Y}$ est un estimateur ***sans biais*** de μ
- Quelle est la variance de $\bar{Y}$?
  - Comment var($\bar{Y}$) dépend-il de n (la formule $1/n$)
- Est-ce que $\bar{Y}$ se rapproche de μ quand n est grand?
  - Loi des grands nombres: $\bar{Y}$ est un estimateur ***convergent*** de μ
- $\bar{Y} - \mu$ suit une distribution en forme de cloche quand n est grand…est-ce généralement le cas?
  - En fait, $\bar{Y} - \mu$ est à peu près normalement distribué pour n grand (Théorème Central de la Limite)

WWW.ECONOMIE-GESTION.COM

Moyenne et variance de la distribution d'échantillonnage de $\bar{Y}$

Cas général – càd, pour Y_i i.i.d. provenant d'une distribution (peu importe laquelle, pas juste la Bernoulli) :

Moyenne: $E(\bar{Y}) = E(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \mu_Y = \mu_Y$

Variance:

$$\begin{aligned}
\mathrm{var}(\bar{Y}) &= E[\bar{Y} - E(\bar{Y})]^2 \\
&= E[\bar{Y} - \mu_Y]^2 \\
&= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) - \mu_Y\right]^2 \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n} (Y_i - \mu_Y)\right]^2
\end{aligned}$$

WWW.ECONOMIE-GESTION.COM

donc

$$\operatorname{var}(\bar{Y}) = E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)\right]^2$$

$$= E\left\{\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)\right]\times\left[\frac{1}{n}\sum_{j=1}^{n}(Y_j - \mu_Y)\right]\right\}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}E\left[(Y_i - \mu_Y)(Y_j - \mu_Y)\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{cov}(Y_i, Y_j)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2$$

$$= \frac{\sigma_Y^2}{n}$$

WWW.ECONOMIE-GESTION.COM

Moyenne et variance de la distribution d'échantillonnage de $\bar{Y}$, suite.

$$\boxed{E(\bar{Y}) = \mu_Y}$$

$$\boxed{\text{var}(\bar{Y}) = \frac{\sigma_Y^2}{n}}$$

Implications:

1. $\bar{Y}$ est un estimateur *sans biais* de μ_Y (càd, $E(\bar{Y}) = \mu_Y$)
2. var($\bar{Y}$) est inversement proportionnelle à n
   - l'écart-type de la distribution d'échantillonnage est proportionnel à $1/\sqrt{n}$
   - Donc l'incertitude d'échantillonnage associée à $\bar{Y}$ est proportionnel à $1/\sqrt{n}$ (échantillons plus grands, moins d'incertitude, mais en fonction de la racine carrée de n)

WWW.ECONOMIE-GESTION.COM

La distribution d'échantillonnage de $\bar{Y}$ lorsque *n* est grand

Pour de petits échantillons, la distribution de $\bar{Y}$ est compliquée, mais si *n* est grand, elle devient simple!

1. Au fur et à mesure que *n* augmente, la distribution de $\bar{Y}$ devient de plus en plus resserrée autour de μ_Y (*Loi des Grands Nombres*)
2. De plus, la distribution de $\bar{Y} - \mu_Y$ devient normale (*Théorème Central de la Limite*)

WWW.ECONOMIE-GESTION.COM

La *Loi des Grands Nombres*:

Un estimateur est ***convergent*** si la probabilité qu'il tombe dans un intervalle autour de la vraie valeur de la population tend vers 1 lorsque la taille de l'échantillon augmente.

Si $(Y_1,\ldots,Y_n)$ sont i.i.d. et que $\sigma_Y^2 < \infty$, lors $\bar{Y}$ est un estimateur convergent de μ_Y, càd,

$$P[|\bar{Y} - \mu_Y| < \varepsilon] \to 1 \text{ lorsque } n \to \infty$$

Ce qu'on peut écrire, $\bar{Y} \xrightarrow{p} \mu_Y$

("$\bar{Y} \xrightarrow{p} \mu_Y$" veut dire "$\bar{Y}$ converge en probabilité vers μ_Y").

(*Les maths*: quand $n \to \infty$, $\text{var}(\bar{Y}) = \dfrac{\sigma_Y^2}{n} \to 0$, ce qui implique que $P[|\bar{Y} - \mu_Y| < \varepsilon] \to 1$.)

WWW.ECONOMIE-GESTION.COM

Le *Théorème Central de la Limite* (CLT):

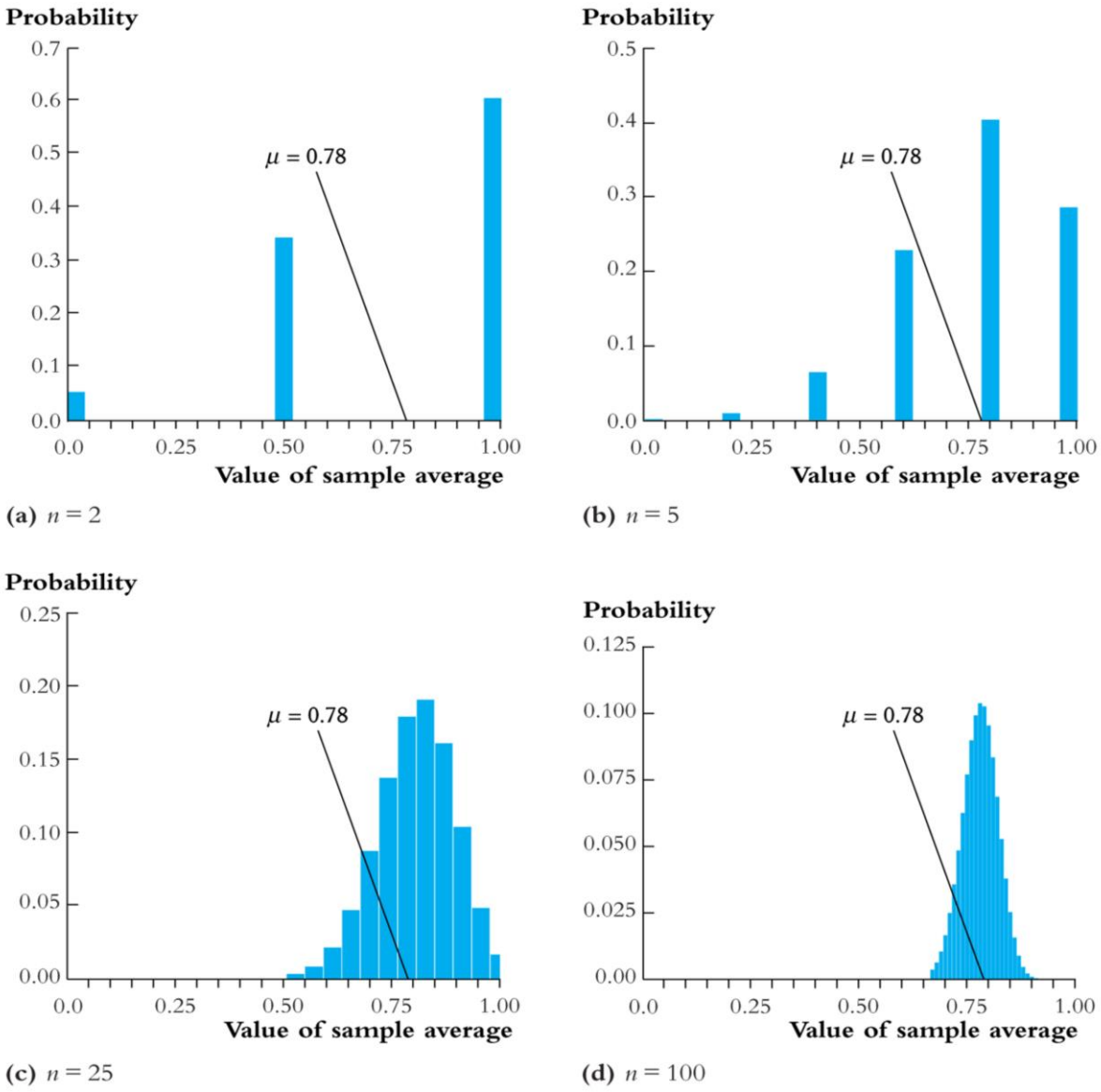
Si $(Y_1,\ldots,Y_n)$ sont i.i.d. et $0 < \sigma_Y^2 < \infty$, alors quand n est grand, la distribution de $\bar{Y}$ est bien approximée par une distribution normale.

- $\bar{Y}$ est approximativement distribuée $N(\mu_Y, \dfrac{\sigma_Y^2}{n})$ ("distribution normale avec moyenne μ_Y et variance σ_Y^2/n")
- $\sqrt{n}(\bar{Y} - \mu_Y)/\sigma_Y$ est approximativement distribué $N(0,1)$ (normale standardisée)
- **Donc, la "moyenne standardisée"** $\bar{Y} = \dfrac{\bar{Y} - E(\bar{Y})}{\sqrt{\mathrm{var}(\bar{Y})}} = \dfrac{\bar{Y} - \mu_Y}{\sigma_Y / \sqrt{n}}$

  est approximativement distribué $N(0,1)$
- **Plus n est grand, plus l'approximation est précise.**

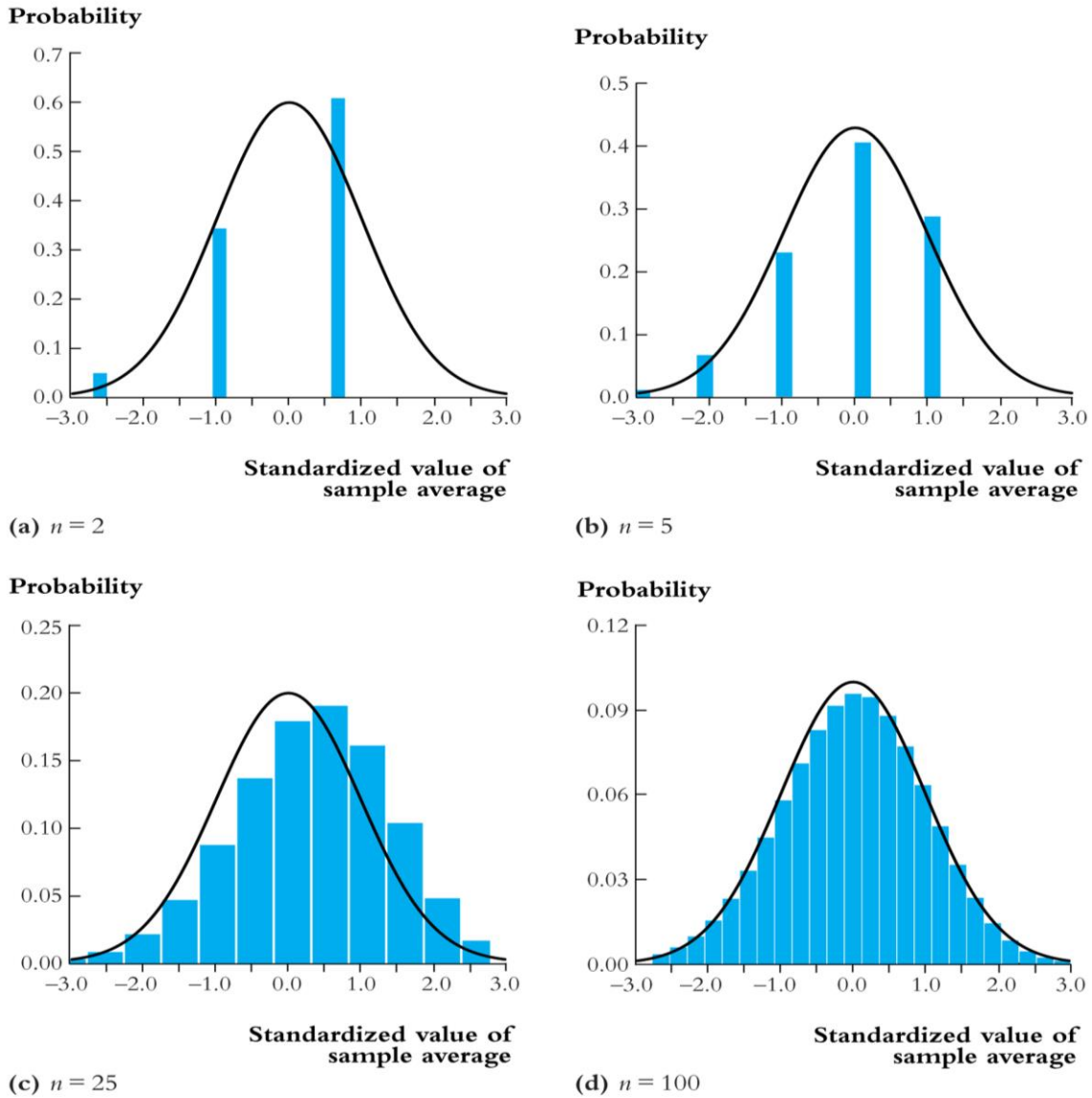
WWW.ECONOMIE-GESTION.COM

Distribution d'échantillonnage de $\bar{Y}$ avec Y Bernoulli (p=0.78)



WWW.ECONOMIE-GESTION.COM

Même exemple: distribution d'échantillonnage de $\dfrac{\bar{Y} - E(\bar{Y})}{\sqrt{\mathrm{var}(\bar{Y})}}$:



WWW.ECONOMIE-GESTION.COM

Résumé: Distribution d'échantillonnage de $\bar{Y}$

Pour $Y_1,\ldots,Y_n$ i.i.d. avec $0 < \sigma_Y^2 < \infty$,

- La distribution d'échantillonnage exacte (avec échantillon fini) de $\bar{Y}$ a une moyenne de μ_Y ("$\bar{Y}$ est un estimateur *sans biais* de μ_Y") et variance σ_Y^2/n
- Outre sa moyenne et sa variance, la distribution exacte de $\bar{Y}$ est compliquée et dépend de la distribution de Y (la distribution de la population)
- Lorsque n est grand, la distribution d'échantillonnage se simplifie:
  - $\boxed{\bar{Y} \xrightarrow{p} \mu_Y}$ (Loi des grands nombres)
  - $\boxed{\dfrac{\bar{Y} - E(\bar{Y})}{\sqrt{\mathrm{var}(\bar{Y})}} \text{ est approximativement } N(0,1)}$ (CLT)

WWW.ECONOMIE-GESTION.COM

(b) Pourquoi utiliser $\bar{Y}$ pour estimer μ_Y?

- $\bar{Y}$ est sans biais: $E(\bar{Y}) = \mu_Y$
- $\bar{Y}$ est convergent: $\bar{Y} \xrightarrow{p} \mu_Y$
- $\bar{Y}$ est l'estimateur des "moindres carrés" de μ_Y; $\bar{Y}$ résout,

$$\min_m \sum_{i=1}^{n} (Y_i - m)^2$$

donc, $\bar{Y}$ minimise la somme des carrés des "résidus" :

$$\frac{d}{dm}\sum_{i=1}^{n}(Y_i - m)^2 = \sum_{i=1}^{n}\frac{d}{dm}(Y_i - m)^2 = 2\sum_{i=1}^{n}(Y_i - m)$$

Mettons la dérivée égale à 0 et appelons $\hat{m}$ la valeur optimale de m:

$$\sum_{i=1}^{n} Y = \sum_{i=1}^{n}\hat{m} = n\hat{m} \text{ ou } \hat{m} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$$

WWW.ECONOMIE-GESTION.COM

Pourquoi utiliser $\bar{Y}$ pour estimer μ_Y, suite

- $\bar{Y}$ a une plus petite variance que tous les autres estimateurs *linéaires sans biais* : considérons l'estimateur $\hat{\mu}_Y = \frac{1}{n}\sum_{i=1}^{n} a_i Y_i$, avec $\{a_i\}$ tels que $\hat{\mu}_Y$ soit sans biais; alors $\text{var}(\bar{Y}) \leq \text{var}(\hat{\mu}_Y)$
- $\bar{Y}$ n'est pas le seul estimateur de μ_Y – pouvez-vous dire dans quelles circonstances serait-il préférable d'utiliser la médiane?

WWW.ECONOMIE-GESTION.COM

1. Le cadre probabiliste de l'inférence statistique
2. Estimation
3. **Tests d'hypothèses**
4. Intervalles de Confiance

Tests d'hypothèse

Le problème des ***tests d'hypothèse*** (pour la moyenne): décider provisoirement si une hypothèse est vraie, ou si au contraire une hypothèse alternative est vraie, sur la base de l'information disponible. Càd, tester

H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) > \mu_{Y,0}$ (unilatéral, >)

H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) < \mu_{Y,0}$ (unilatéral, <)

H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) \neq \mu_{Y,0}$ (bilatéral)

WWW.ECONOMIE-GESTION.COM

Un peu de terminologie pour les tests d'une hypothèse statistique:

p-valeur = probabilité de tirer une statistique (p. ex. $\bar{Y}$) au moins aussi défavorable à l'hypothèse nulle que la valeur calculée dans l'échantillon, mais en supposant vraie l'hypothèse nulle.

Le ***niveau de signification*** d'un test est la probabilité spécifiée à l'avance de se tromper en rejetant l'hypothèse nulle alors qu'elle est vraie.

Calcul de la p-valeur basée sur $\bar{Y}$:

$$\boxed{p\text{-}valeur = \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|]}$$

où $\bar{Y}^{act}$ est la valeur de $\bar{Y}$ observée (non-stochastique)

WWW.ECONOMIE-GESTION.COM

Calcul de la p-valeur, suite.

- Pour calculer la *p*-valeur, il faut connaître la distribution d'échantillonnage de $\bar{Y}$, qui est compliquée quand *n* est petit.
- Si *n* est grand, on peut utiliser l'approximation normale (CLT):

p-valeur= $\Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|]$,

$$= \Pr_{H_0}[|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}| > |\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}|]$$

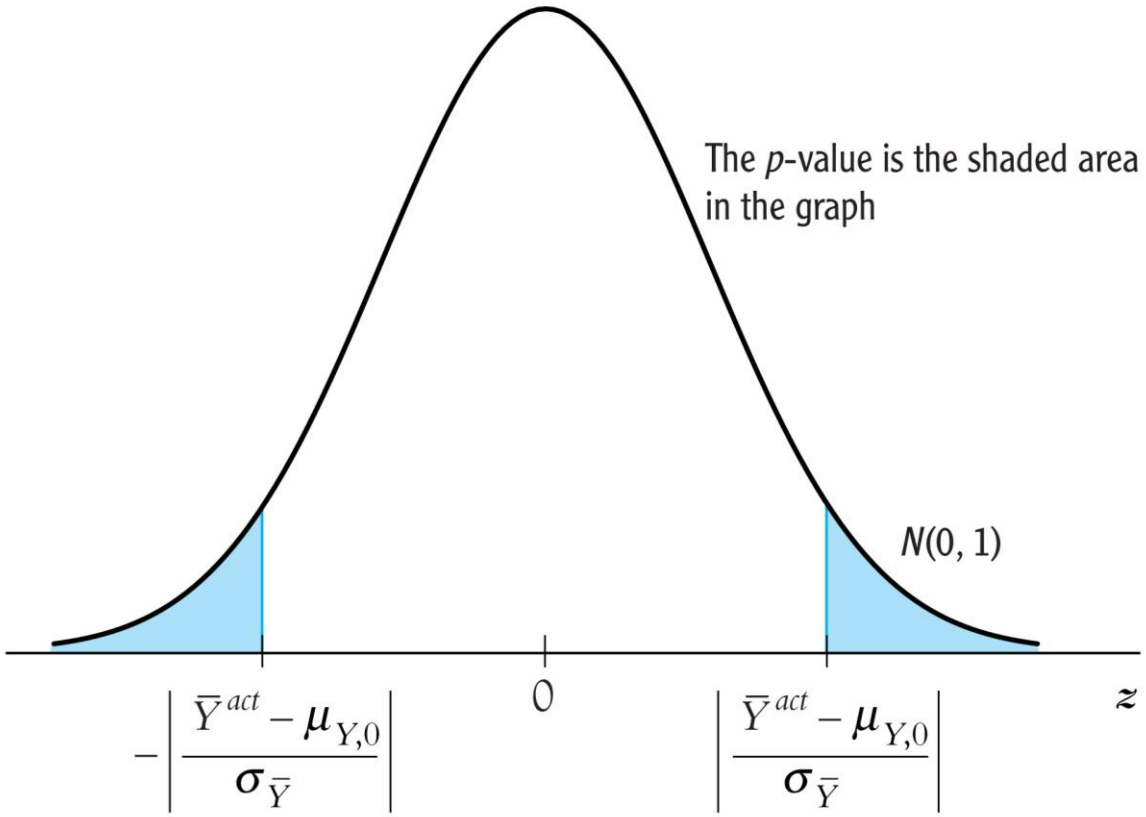
$$= \Pr_{H_0}[|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}| > |\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}|]$$

$\cong$ probabilité sous les queues droite et gauche de la $N(0,1)$

avec $\sigma_{\bar{Y}}$ = écart-type de la distribution de $\bar{Y}$ = $\sigma_Y/\sqrt{n}$.

WWW.ECONOMIE-GESTION.COM

Calcul de la p-valeur lorsque σ_Y est connue:



- Pour des valeurs grandes de *n*, *p-valeur*= probabilité qu'une v.a. $N(0,1)$ soit en-dehors de l'intervalle $|(\bar{Y}^{act}-\mu_{Y,0})/\sigma_{\bar{Y}}|$
- En pratique, $\sigma_{\bar{Y}}$ est inconnue et doit être estimée

WWW.ECONOMIE-GESTION.COM

Estimateur de la variance de Y:

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \text{“variance de Y dans l'échantillon”}$$

Fait:

Si $(Y_1,\ldots,Y_n)$ sont i.i.d. et $E(Y^4) < \infty$, alors $s_Y^2 \xrightarrow{p} \sigma_Y^2$

Pourquoi la loi des grands nombre s'applique-t-elle?

- Parce que s_Y^2 est une moyenne dans l'échantillon
- Note technique: nous faisons l'hypothèse que $E(Y^4) < \infty$ parce qu'ici la moyenne n'est pas de Y_i, mais du carré de Y_i.

WWW.ECONOMIE-GESTION.COM

Calcul de la p-valeur avec σ_Y^2 estimé:

$$p\text{-}valeur = \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|],$$

$$= \Pr_{H_0}[|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}| > |\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}|]$$

$$\cong \Pr_{H_0}[|\frac{\bar{Y} - \mu_{Y,0}}{s_Y / \sqrt{n}}| > |\frac{\bar{Y}^{act} - \mu_{Y,0}}{s_Y / \sqrt{n}}|] \quad (n \text{ grand})$$

alors

$$\boxed{p\text{-}valeur = \Pr_{H_0}[|t| > |t^{act}|]} \quad \textbf{(}\sigma_Y^2 \textbf{ estimé)}$$

$\cong$ probabilité sous la normale en-dehors de $|t^{act}|$

alors $t = \dfrac{\bar{Y} - \mu_{Y,0}}{s_Y / \sqrt{n}}$ (la t-stat habituelle)

WWW.ECONOMIE-GESTION.COM

Quel est le lien entre la *p-valeur* et le niveau de signification?

Le niveau de signification est spécifié à l'avance. Par exemple, si ce niveau est de 5%,

- On rejette l'hypothèse nulle si $|t| \geq 1.96$.
- De façon équivalente, on rejette si $p \leq 0.05$.
- La *p-valeur* est parfois appelée le ***niveau marginal de signification***.
- Souvent, il est préférable de communiquer la *p-valeur* plutôt que juste si un test rejette ou non – la *p-valeur* contient plus d'information que juste une réponse "oui/non" sur le résultat du test.

WWW.ECONOMIE-GESTION.COM

Digression: la distribution t de Student

Si Y_i, $i = 1,\ldots, n$ est i.i.d. $N(\mu_Y,\sigma_Y^2)$, alors la t-stat suit une distribution de Student avec $n - 1$ degrés de liberté.

Les valeurs critiques de la distribution t de Student sont tabulées dans tous les livres de statistique. Vous vous souvenez de la recette?

1. Calculez la t-stat
2. Calculez les degrés de liberté, qui sont égaux à $n - 1$
3. Chercher la valeur critique à 5% dans la table
4. Si la t-stat est supérieure (en valeur absolue) à la valeur critique, on rejette l'hypothèse nulle.

WWW.ECONOMIE-GESTION.COM

Commentaires sur la recette et la distribution *t* de Student

1. La théorie de la distribution *t* est l'un des premiers triomphes de la statistique mathématique. Il est étonnant: si *Y* est i.i.d. normale, alors on connaît exactement, la distribution en échantillon fini de la statistique de t –c'est la distribution t de Student. Donc, on peut construire des intervalles de confiance (en utilisant la valeur critique du t de Student) qui ont exactement le taux de couverture correct, quelle que soit la taille de l'échantillon. Ce résultat était vraiment utile à une époque où l'on ne disposait pas d'ordinateurs, où la collecte de données était très onéreuse, et l'on se contentait d'une douzaine d'observations. C'est aussi du point de vue conceptuel un beau résultat, et le calcul est élégant - ce qui est probablement pourquoi les profs de stat aiment enseigner la distribution t. Mais...

WWW.ECONOMIE-GESTION.COM

Commentaires sur la distribution t de Student, suite

2. Si la taille de l'échantillon est modérée (quelques dizaines) ou grande (des centaines ou plus), la différence entre la *distribution t de Student* et les valeurs critiques de la N(0,1) est négligeable. Voici les valeurs critiques à 5% pour des tests bilatéraux:

degrés de liberté ($n - 1$)	valeurs critiques à 5% de la t
10	2.23
20	2.09
30	2.04
60	2.00
∞	1.96

WWW.ECONOMIE-GESTION.COM

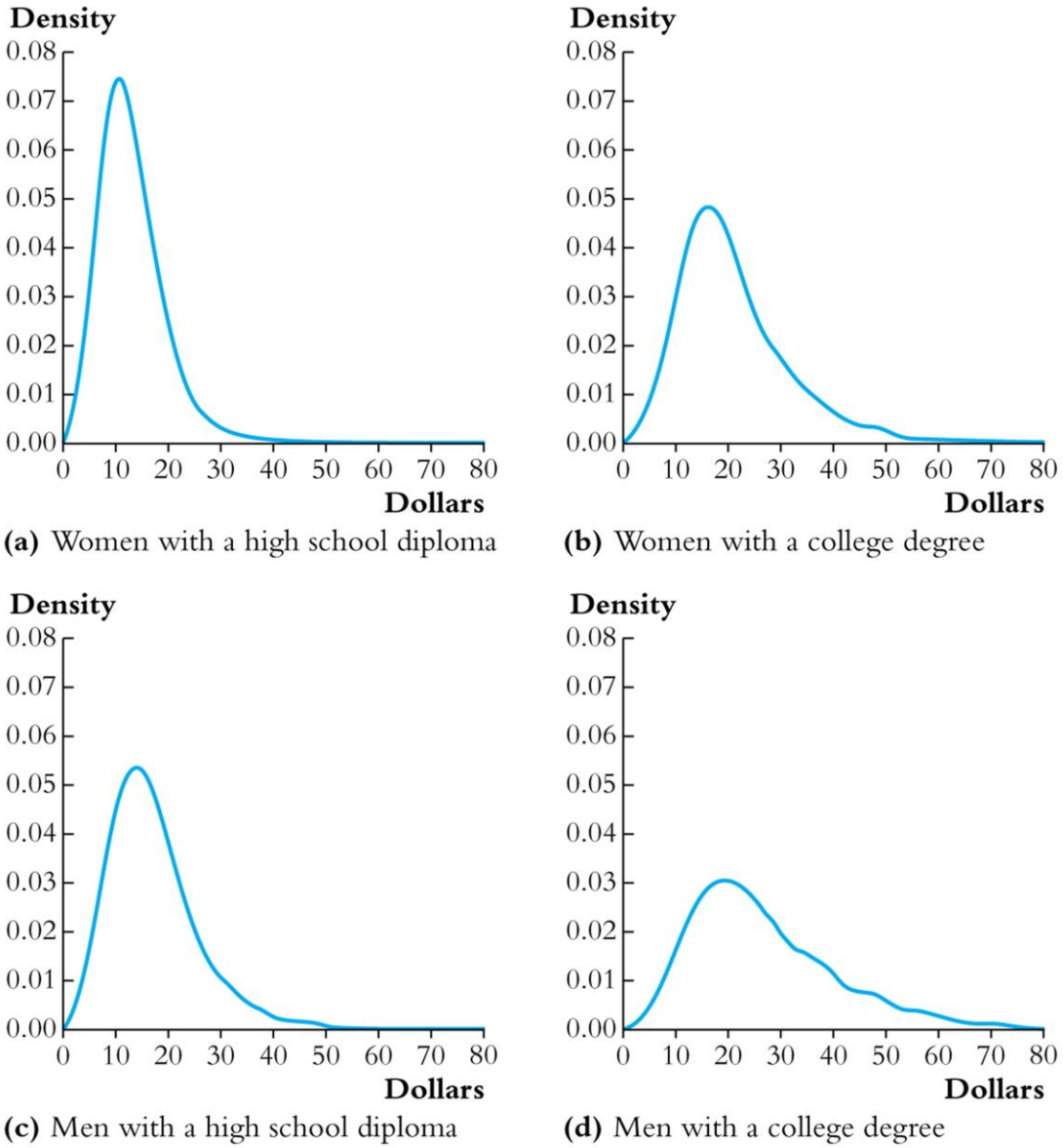
Commentaires sur la distribution t de Student, suite

- Donc, la distribution t de Student n'est vraiment pertinente que lorsque la taille de l'échantillon est très faible, mais dans ce cas, pour qu'il soit correct, vous devez être sûr que la distribution Y de la population est bien normale. Avec les données économiques, l'hypothèse de normalité est rarement crédible. Voici les distributions de quelques variables économiques.
- Pensez-vous que les revenus sont normalement distribués?
- Supposons que vous avez un échantillon de n = 10 observations de l'une de ces distributions - vous sentiriez à l'aise avec la distribution t de Student?

WWW.ECONOMIE-GESTION.COM

FIGURE 2.4 **Conditional Distribution of Average Hourly Earnings of U.S. Full-Time Workers in 2004, Given Education Level and Gender**

The four distributions of earnings are for women and men, for those with only a high school diploma (a and c) and those whose highest degree is from a four-year college (b and d).



WWW.ECONOMIE-GESTION.COM

Commentaires sur la distribution t de Student, suite.

3. Considérons la *t*-stat pour l'hypothèse que deux moyennes (groupes *s*, *l*) sont égales:

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{\bar{Y}_s - \bar{Y}_l}{SE(\bar{Y}_s - \bar{Y}_l)}$$

Même si la distribution de la population de Y dans les deux groupes est normale, cette statistique ne suit pas une distribution *t* de Student!

Il existe une statistique distribuée normalement pour tester cette hypothèse, c'est la statistique de *t* avec "variance groupée" - mais elle n'est valable que si les variances des deux distributions normales sont identiques dans les deux groupes. Pensez-vous que ce soit vrai, par exemple, pour les salaires des femmes des hommes ?

WWW.ECONOMIE-GESTION.COM

La distribution t de Student– Résumé

- L'hypothèse que Y est distribuée $N(\mu_Y, \sigma_Y^2)$ est rarement crédible en pratique (Revenu? Nombre d'enfants?)
- Quand $n > 30$, les distributions de t et $N(0,1)$ sont très proches (quand n devient grand, la distribution t_{n-1} converge vers la $N(0,1)$)
- La distribution de t date de l'époque où les échantillons étaient petits et les ordinateurs n'existaient pas.
- Pour des raisons historiques, les logiciels statistiques utilisent encore la distribution de t pour calculer les p-valeurs, mais c'est inutiles quand les échantillons sont de taille modérée ou grande.
- Pour cette raison, dans ce cours, nous allons nous focaliser sur des approximations pour n grand, qui découlent du CLT

WWW.ECONOMIE-GESTION.COM

1. Le cadre probabiliste de l'inférence statistique
2. Estimation
3. Tests d'hypothèse
4. **Intervalles de Confiance**

Intervalles de Confiance

Un ***intervalle de confiance*** à 95% de μ_Y est un intervalle qui contient la vraie valeur de μ_Y dans 95% des échantillons répétés.

Digression: Qu'est-ce qui est stochastique ici? Les valeurs de $Y_1,\ldots,Y_n$ et donc toute fonction calculée à partir de ces valeurs– y compris l'intervalle de confiance. Cet intervalle de confiance sera différent d'un échantillon à l'autre. Le paramètre, μ_Y, de la population n'est pas aléatoire; il est juste inconnu.

WWW.ECONOMIE-GESTION.COM

Intervalle de confiances, suite.

Un intervalle de confiance à 95% peut toujours être construit comme l'ensemble de valeurs de μ_Y qui ne sont pas rejetées par un test d'hypothèse au niveau de signification de 5%.

$$\{\mu_Y: \left|\frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}}\right| \leq 1.96\} = \{\mu_Y: -1.96 \leq \frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}} \leq 1.96\}$$

$$= \{\mu_Y: -1.96\frac{s_Y}{\sqrt{n}} \leq \bar{Y} - \mu_Y \leq 1.96\frac{s_Y}{\sqrt{n}}\}$$

$$= \{\mu_Y \in (\bar{Y} - 1.96\frac{s_Y}{\sqrt{n}}, \bar{Y} + 1.96\frac{s_Y}{\sqrt{n}})\}$$

Cet intervalle de confiance est basé sur le résultat avec n grand que $\bar{Y}$ *est approximativement distribué comme un normale et que* $s_Y^2 \xrightarrow{p} \sigma_Y^2$.

WWW.ECONOMIE-GESTION.COM

Résumé:

A partir des hypothèses suivantes:

(1) Échantillonnage aléatoire simple à partir d'une population, càd que les

$\{Y_i, i = 1,\ldots,n\}$ sont i.i.d.

(2) $0 < E(Y^4) < \infty$

Nous avons développé, pour des échantillons grands (n grand):

- Une théorie de l'estimation (distribution d'échantillonnage de $\bar{Y}$)
- Une théorie des tests d'hypothèse (distribution avec n grand des t-stat et calcul des p-valeurs)
- Une théorie des intervalles de confiance (construits en inversant la statistique du test)

Est-ce que les hypothèses (1) & (2) sont crédibles en pratique? **Oui**

WWW.ECONOMIE-GESTION.COM

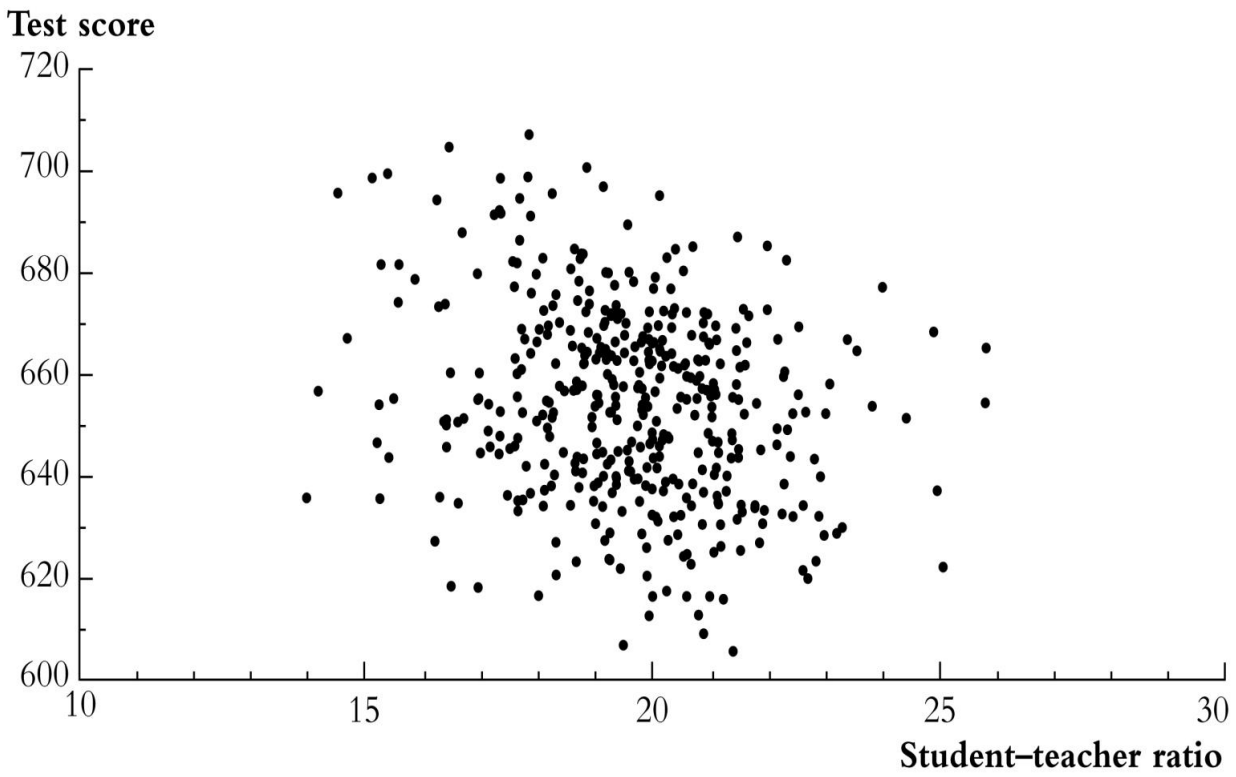
Revenons à la question de politique économique initiale:

Quel est l'effet d'une réduction de la taille des classes sur les résultats scolaires ?

Avons-nous répondu à la question?

FIGURE 4.2 Scatterplot of Test Score vs. Student–Teacher Ratio (California School District Data)

Data from 420 California school districts. There is a weak negative relationship between the student–teacher ratio and test scores: The sample correlation is −0.23.



WWW.ECONOMIE-GESTION.COM