

Statistiques pour les sciences (MAT-4681)

Arthur Charpentier

05 - Échantillon, Observations & Expériences

été 2022

Échantillonnage I

Population (N)

La population est constituée de l'ensemble des unités auxquelles les résultats de l'enquête s'appliqueront (taille N).

Base d'échantillonnage (n)

La base d'échantillonnage ou base de sondage est constituée par la liste des unités d'échantillonnage (liste matérielle ou conceptuelle), c'est-à-dire liste des unités à partir de laquelle se fera la sélection (taille n). Cette liste doit constituer la meilleure approximation possible de la population

Échantillonnage II

► L'échantillonnage probabiliste

Échantillon probabiliste

Chaque unité doit avoir une probabilité connue d'être choisie. Cette probabilité ne peut pas être nulle (mais elle n'est pas nécessairement égale pour toutes les unités).

Tirage aléatoire simple

On sélectionne les unités au hasard, uniformément, sans remise.

```
1 > sample(1:10000, size = 20, replace = FALSE)
2 [1] 4226 8803 7632 3542 4525 2191 5699 8823 6052 7365
3 [11] 5649 1257 767 7587 8477 9826 481 5916 8707 7441
```

Échantillonnage III



Pour le réaliser, il faut avoir une liste de la population (la base d'échantillonnage)

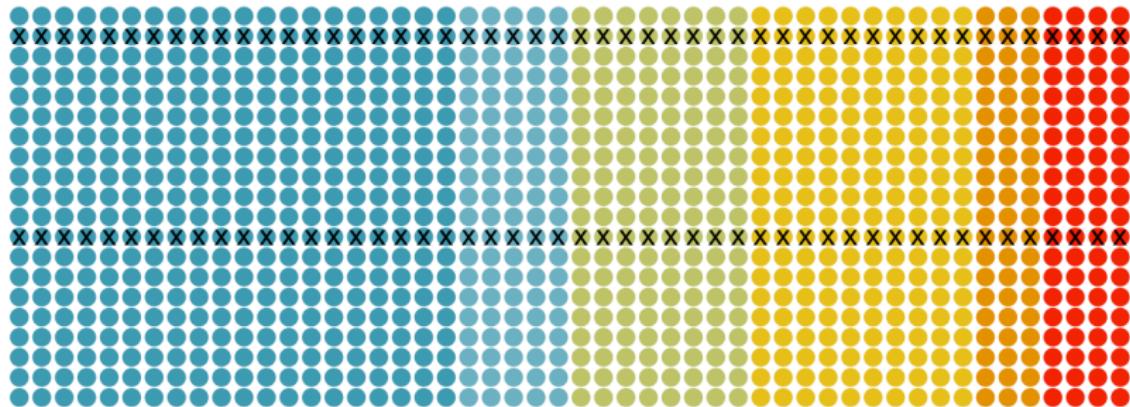
Échantillonnage IV

Tirage aléatoire systématique

Il s'agit ici de tirer seulement la première unité de la liste au hasard, et de prendre ensuite les unités à un intervalle prédéterminé.

```
1 > sample(1:10000, size = 1) + (0:19)*41
2 [1] 5222 5263 5304 5345 5386 5427 5468 5509 5550 5591
3 [11] 5632 5673 5714 5755 5796 5837 5878 5919 5960 6001
```

Échantillonnage V

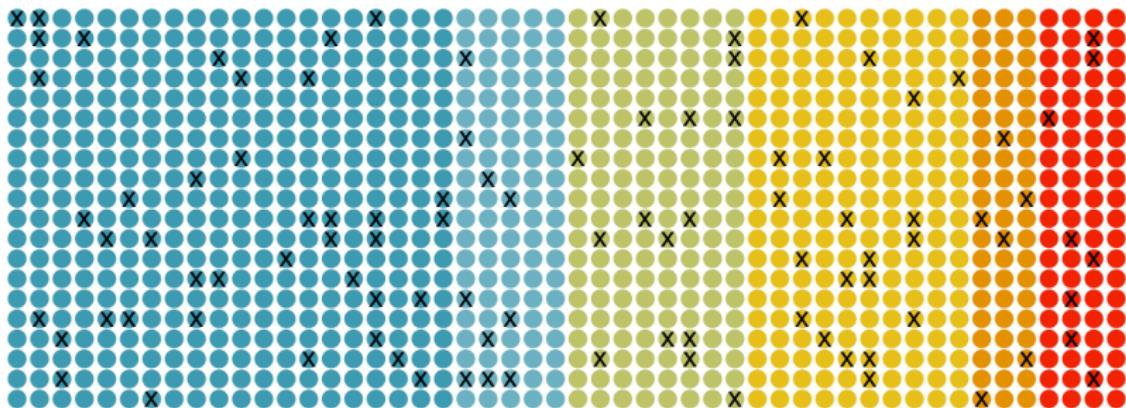


Échantillon stratifié

On parle d'échantillon stratifié lorsque la fraction de sélection est fonction de certaines caractéristiques de la population (sexe, région, statut, âge, etc.)

Échantillonnage VI

On stratifie pour permettre que toutes les catégories de la population qui nous intéressent soient représentées en nombre suffisant.



Les échantillons stratifiés peuvent être proportionnels, c'est-à-dire que l'on tire proportionnellement le même nombre d'unités dans chaque strate, ou pas.

Échantillonnage VII

Échantillon par grappes

Dans l'échantillonnage par grappes, on sélectionnera un certain nombre d'unités puis les unités contiguës (grappes).

► L'échantillonnage non-probabiliste

L'échantillon raisonné

Il s'agit ici de choisir des unités (quartiers, îlots, écoles,...) en fonction de certaines caractéristiques. L'échantillonnage à l'intérieur de chacune des grandes unités se fait ensuite au hasard

Échantillonnage VIII

L'échantillon par quota

Il s'agit ici de déterminer le nombre de personnes possédant chaque caractéristique de base que l'on veut dans l'échantillon et d'arrêter de recueillir les données dès que ce nombre (le quota) est atteint.

L'échantillon de volontaires

L'échantillonnage de volontaires ne peut être considéré comme représentatif d'une population. On l'utilise uniquement lorsque l'on peut prétendre que les phénomènes étudiés sont intra-individuels et universels, lorsque l'on étudie des processus pour eux-mêmes

Échantillonnage IX

Pour un échantillon dans une population de taille infinie ($N > 20 \times n$).

La marge d'erreur, c'est la précision du résultat obtenu étant donné le seuil de confiance que l'on est prêt à accepter. La marge d'erreur (absolue) est alors égale à

$$z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

```
1 > qnorm(.975)*sqrt(.5*(1-.5)/1000)
2 [1] 0.03098975
```

soit 3.1% (intervalle à 95%) pour $n = 1,000$ et $p \sim 50\%$

Échantillonnage X

Pour un échantillon d'une population finie, une correction peut s'imposer

$$z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$$

```
1 > qnorm(.975)*sqrt(.5*(1-.5)/1000)*sqrt(5000/5999)
2 [1] 0.028292
```

soit 2.8% (intervalle à 95%) pour $n = 1,000$, $p \sim 50\%$, et $N = 6,000$.

On peut inverser ces formules pour déterminer la taille d'échantillonnage, en fonction de la marge d'erreur (absolue)

$$n = \frac{z_{1-\alpha/2}^2 \times p(1-p)}{\text{marge d'erreur}^2}$$

Échantillonnage XI

ou

$$n = \frac{p(1-p) + \frac{\text{marge d'erreur}^2}{z_{1-\alpha/2}^2}}{\frac{\text{marge d'erreur}^2}{z_{1-\alpha/2}^2} + \frac{p(1-p)}{N}}$$

Observations ou Expériences ? I

cf [Une courte histoire des expériences randomisées](#),

Polio hit the U.S. in 1916, caused hundreds of thousands of fatalities over 40 years. Jonas Salk developed a vaccine, ready to test it in 1954



(images: <https://csnbbs.com/>)

Observations ou Expériences ? II

- ▶ Incidence of disease vary from year to year (1954 vs. 1953)
- ▶ used **controlled experiment**, treatment group vs. control group
- ▶ initial design : select 2 million children (targeted school districts, high risk), vulnerable age (grades 1,2,3)
- ▶ grade 2 gets the vaccine, grades 1 & 3 are the controls
- ▶ (problem parental consent, correlated with higher income)
- ▶ alternative design : randomized controlled double-blind
- ▶ control group in the same population as treatment group
- ▶ random allocation, children, parents and doctors should not know the group

Observations ou Expériences ? III

	grades 1,2,3 size	rate*		randomized size	rate*
treatment	221,998	25	treatment	200,745	28
control	725,173	54	control	201,229	71
no consent	123,605	44	no consent	338,778	46

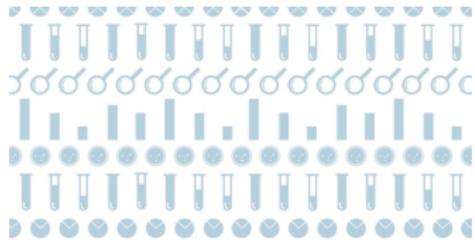
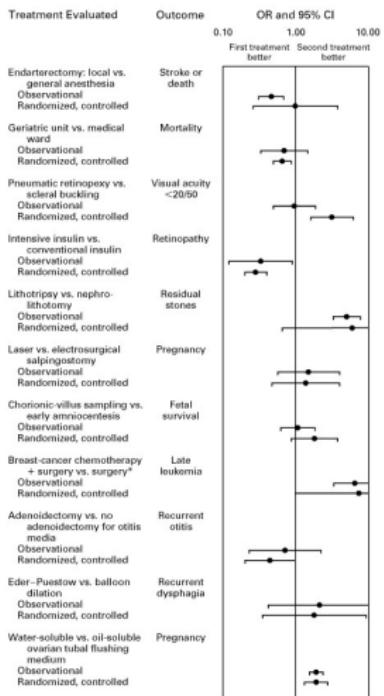
* rate per 100,000

Source : Meier (1972)

- ▶ Controlled experiment: investigators decide who is in the treatment group and who is in the control group.
- ▶ Observational study: the subjects assign themselves to these two groups. The investigators just watch.
- ▶ In some cases, randomized experiments are impossible

Observations ou Expériences ? IV

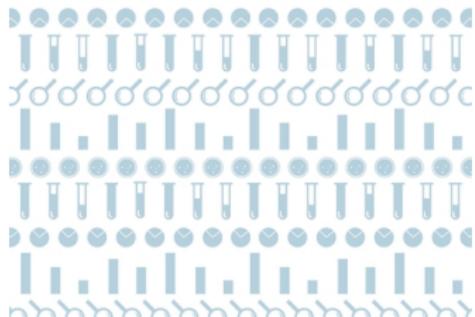
A Comparison of Observational Studies and Randomized, Controlled Trials (in 2000), and Randomised controlled trials the gold standard for effectiveness research (in 2018)



Observation & Experiment

An Introduction to Causal Inference

PAUL R. ROSENBAUM



Graduate admissions data, Berkeley, 1973

Graduate admissions data from Berkeley, 1973

- ▶ men : 8442 applications, 44% admission rate
- ▶ women : 4321 applications, 35% admission rate

Discrimination towards women ?

		A	B	C	D	E	F
M	applied	825	560	325	417	191	373
	admitted	62%	63%	37%	33%	28%	6%
W	applied	108	25	593	375	393	341
	admitted	82%	68%	34%	35%	24%	7%

see Bickel *et al.* (1975, Sex bias in graduate admissions)

(Fake) Hospital Data

	hosp. A	hosp. B
total	1000	1000
survivors	800	900
deads	200	100
rate (%)	80%	90%

	healthy	
	hosp. A	hosp. B
total	600	900
survivors	590	870
deads	10	30
rate (%)	98%	97%

	sick	
	hosp. A	hosp. B
total	400	100
survivors	210	30
deads	190	70
rate (%)	53%	30%

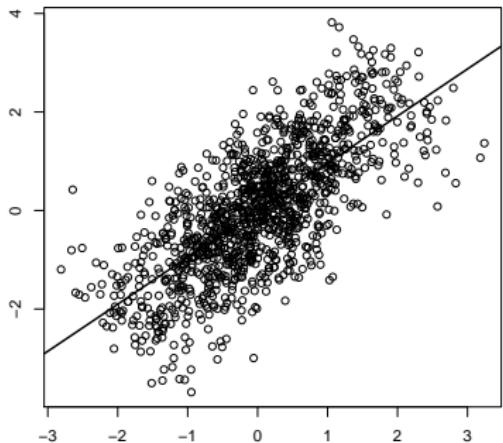
Mathematics of Simpson's Paradox

Heuristically, it is possible to have

$$\frac{a}{A} \leq \frac{b}{B} \text{ and } \frac{c}{C} \leq \frac{d}{D}$$

and at the same time

$$\frac{a+c}{A+C} \geq \frac{b+d}{B+D}$$



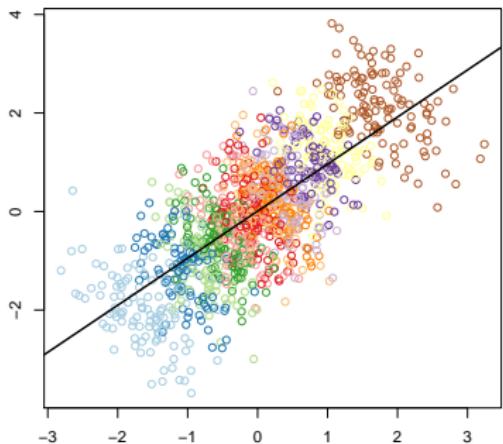
Mathematics of Simpson's Paradox

Heuristically, it is possible to have

$$\frac{a}{A} \leq \frac{b}{B} \text{ and } \frac{c}{C} \leq \frac{d}{D}$$

and at the same time

$$\frac{a+c}{A+C} \geq \frac{b+d}{B+D}$$



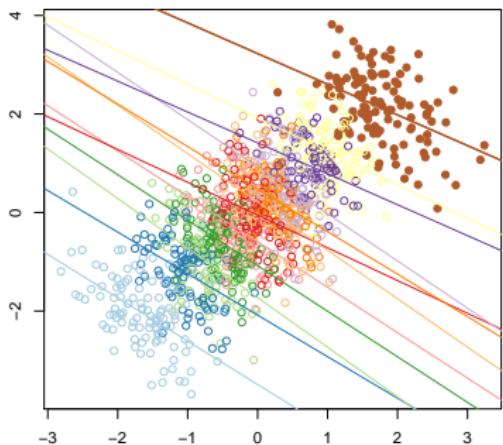
Mathematics of Simpson's Paradox

Heuristically, it is possible to have

$$\frac{a}{A} \leq \frac{b}{B} \text{ and } \frac{c}{C} \leq \frac{d}{D}$$

and at the same time

$$\frac{a+c}{A+C} \geq \frac{b+d}{B+D}$$



Ecological Fallacy

An ecological fallacy is a formal fallacy in the interpretation of statistical data that occurs when inferences about the nature of individuals are deduced from inferences about the group to which those individuals belong, via [wikipedia](#))

See Robinson's [Ecological Correlations and the Behavior of Individuals](#) the individual correlation depends upon the internal frequencies of the within-areas individual correlations, while the ecological correlation depends upon the marginal frequencies of the within-areas individual correlation

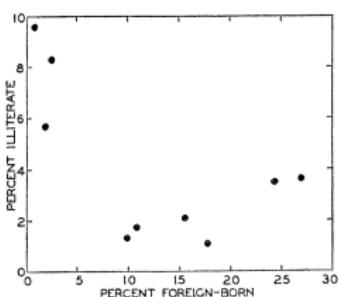
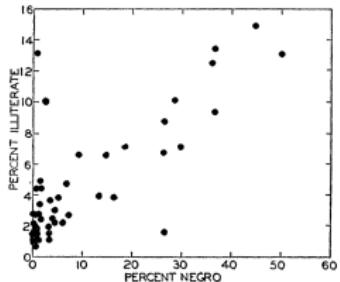
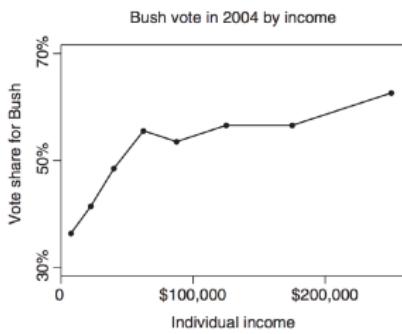
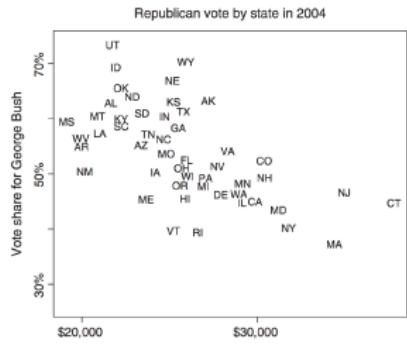
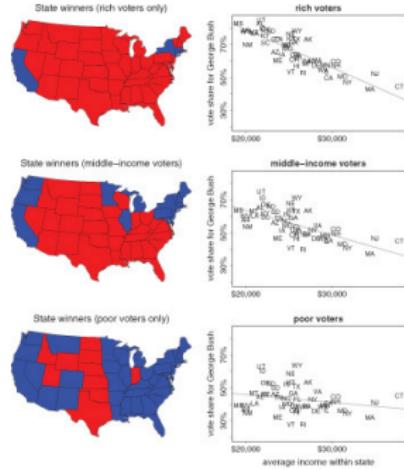
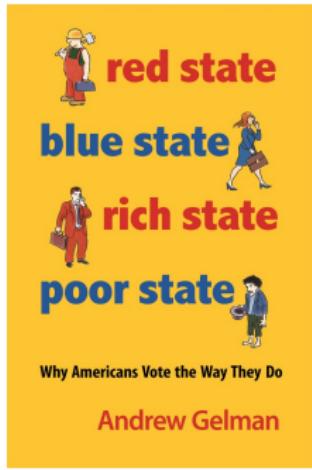


TABLE 3. THE INDIVIDUAL CORRELATION BETWEEN NATIVITY AND ILLITERACY FOR THE UNITED STATES, 1930
(for the population 10 years old and over)

	Foreign Born	Native Born	Total
Illiterate	1,304	2,614	3,918
Literate	11,913	81,441	93,354
Total	13,217	84,055	97,272

Ecological Fallacy

Very important concept in political science



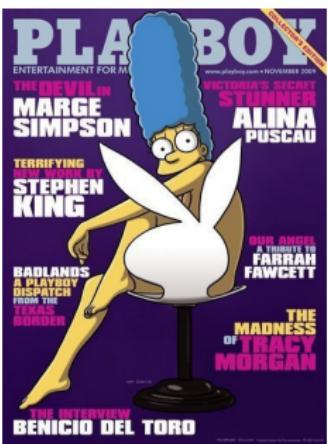
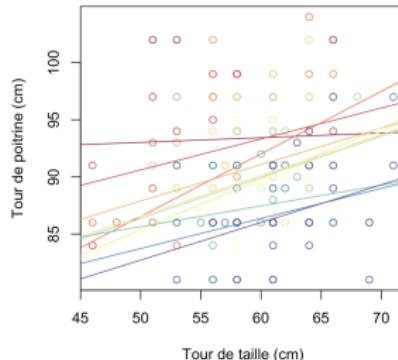
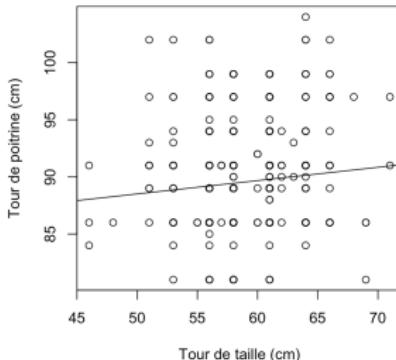
Gelman's Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do)

Playboy: Individual vs. Temporal Data

Are **bust/chest** and **waist** correlated measures ?

Dataset $n = 659$ observations (~ 55 years) of Playboy's playmate (inspired by **Shapely centre-folds**. Are women changing or is Playboy?).

- ▶ x_i : waist (cm)
- ▶ y_i : bust (cm)
- ▶ t_i : date



Playboy: Individual vs. Temporal Data

over 55 years
 $\text{cor}(x_i, y_i) \approx 0.1$

underestimation !

