

Connaître la (bio)statistique pour reconnaître les paradoxes: L'exemple de la COVID-19



Janie Coulombe

Professeure Adjointe
Université de Montréal

Dans le cadre de l'école d'été de
l'Université de Montréal
Le 17 juin 2022

Qui suis-je?



Galaxie M81, Crédit photo Nicolas Michaud 2022

Plan de l'atelier

9h30 à 11h00

Introduction

Etre confus par le paradoxe de Simpson

Le paradoxe de l'amitié dans l'étude de la propagation du virus

Le paradoxe de Berkson: Gare à la stratification!

11h à 11h30

Pause café

11h30 à 12h30

Atelier pratique: Programmation en R

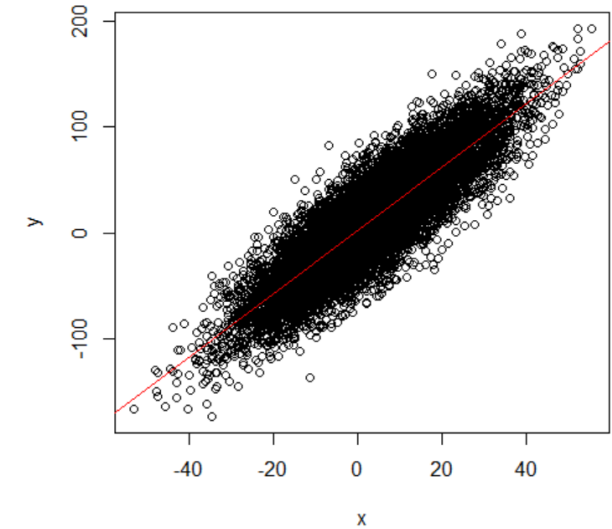
La statistique

- ❑ En statistique, on utilise les données provenant d'un échantillon de la population d'intérêt pour inférer sur des paramètres de la population.
- ❑ Une population d'intérêt: Tous les patients de l'Hôpital Jean-Talon.
- ❑ Exemple d'inférence:



- J'ai accès à un échantillon aléatoire de tailles (en mètres) de patients de l'Hôpital Jean-Talon à Montréal ($n=60$ individus).
- Je souhaite estimer la taille moyenne de tous les patients de l'Hôpital.

La statistique



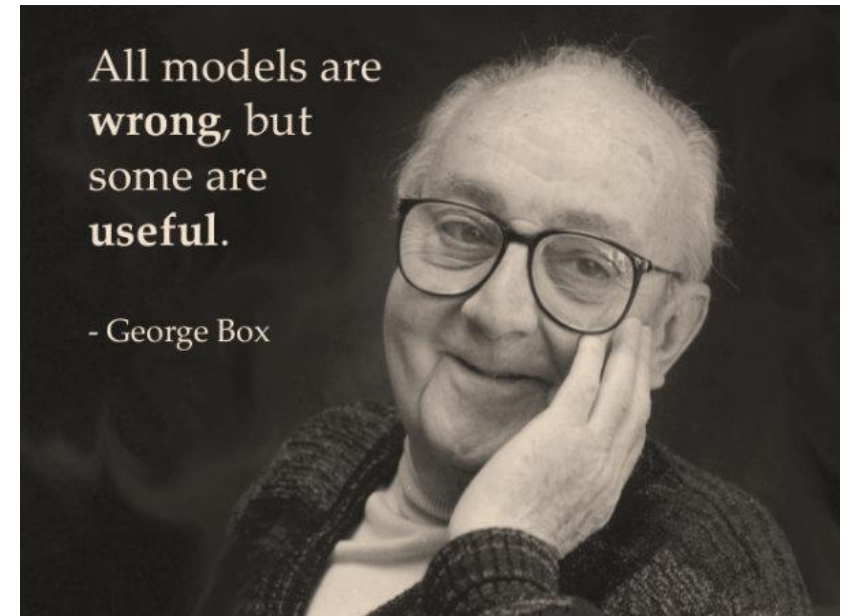
- ❑ On s'intéresse à l'estimation de quantités "résumant" la population.
- ❑ Exemples de quantités d'intérêt:

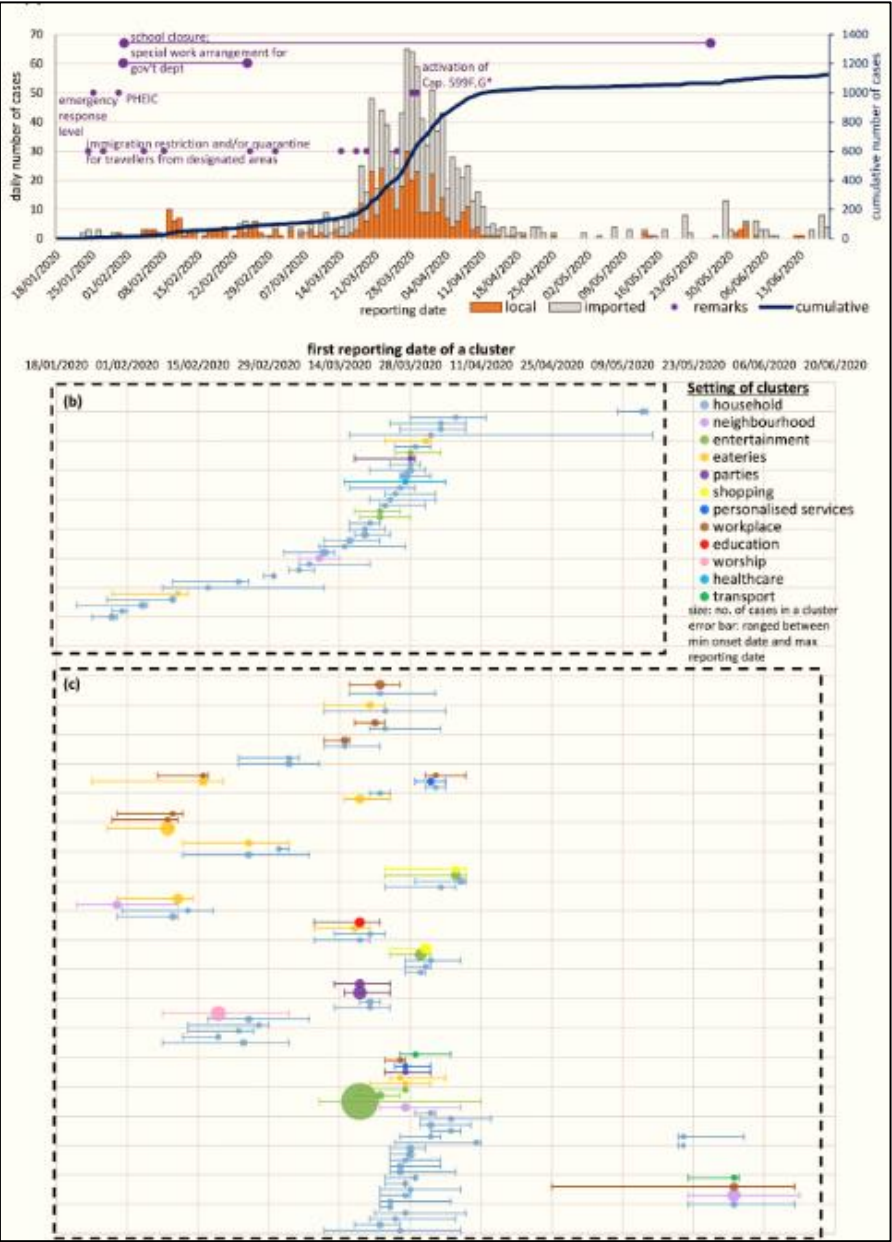
- **Moyenne d'une variable (ex., taille, âge, ou poids moyen)**
- **Proportion, taux, fréquence (ex., taux de contamination à la COVID19)**
- **Associations (ex., l'association entre l'âge et la prise de poids)**
- **Effet causal (ex., l'effet causal de faire de l'exercice sur la pression sanguine)**

$$\bar{X} = \frac{(X_1 + X_2 + \cdots + X_n)}{(n)} = \frac{1 \sum X_j}{n}$$

La statistique

- ❑ L'estimation de ces quantités populationnelles peut vite devenir compliquée.
- ❑ L'estimation se base souvent sur des modèles qu'on a nous-même choisis.
- ❑ On souhaite aussi quantifier l'incertitude de notre estimation.
- ❑ Estimer peut aussi vouloir dire modéliser (page suivante).





Source: <https://www.youtube.com/watch?v=iPPGfEA2s2M>

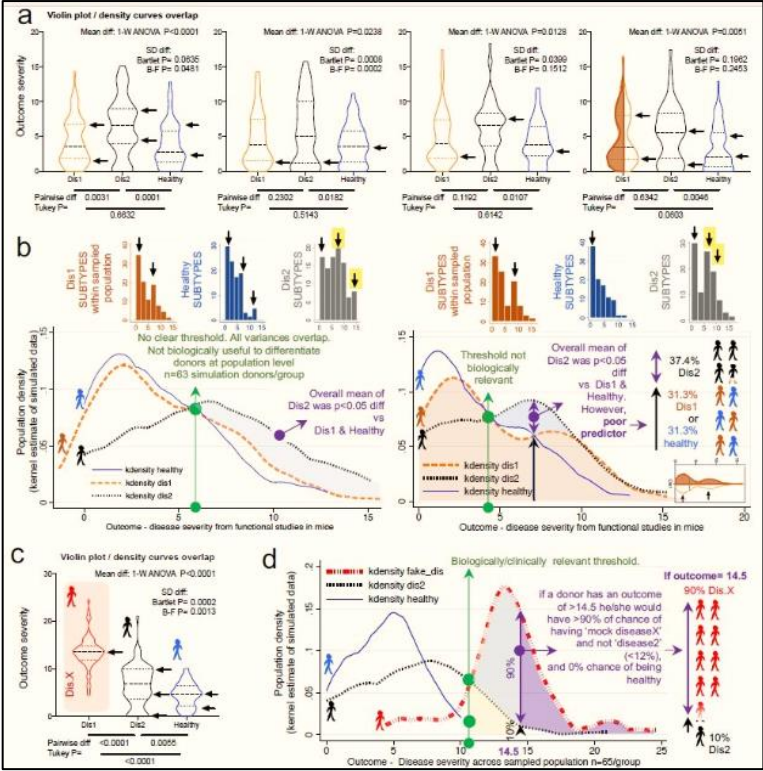
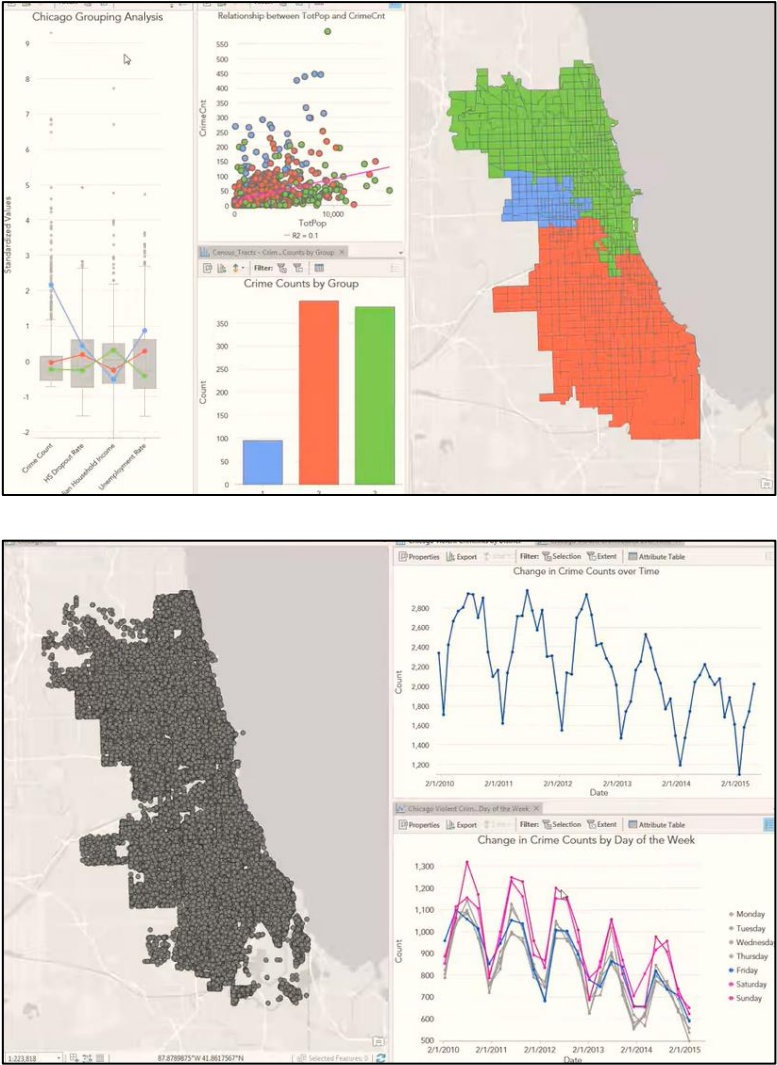


Figure 5 dans Basson et al. (2020), Patterns of analytical irreproducibility in multimodal diseases

Figure 1, Wong, Lee, Kwan and Yeoh (2020), The Lancet

La biostatistique

❑ Une branche de la statistique où on se concentre surtout sur des problèmes en lien avec la santé/biologie ou l'environnement.

❑ Exemples:

- **Efficacité des traitements et comparaison entre deux traitements**
- **Prédiction de maladie (oui/non) ou du temps jusqu'à la maladie**
- **Les relation temporelles ou spatiales entre les maladies et d'autres facteurs**
- **L'effet causal d'une intervention**
- **Le meilleur traitement à donner à quelqu'un, selon ses caractéristiques**

La biostatistique

□ Des amis et collègues qui travaillent en biostatistique:

- **Consultant.e dans un hôpital**
- **Analyste dans un centre de recherche**
- **Programmeur.se pour une compagnie pharmaceutique**
- **Professeur.e en statistique ou en biostatistique**



Les paradoxes

❑ Question d'importance avant de débiter:

Qu'est-ce qu'un paradoxe?

Définition tirée du Larousse:

1. Opinion contraire aux vues communément admises: *Soutenir des paradoxes.*
2. Être, chose ou fait qui paraissent défier la logique parce qu'ils présentent des aspects **contradictaires**: *Cette victoire du plus faible, c'est un paradoxe.*
3. En logique, synonyme de antinomie.

1

Le paradoxe de Simpson

- ❑ Une histoire de 1973 à l'Université de Berkeley Californie
- ❑ Le processus d'admission aux études graduées
- ❑ Données marginales:

| Applicants | Admis | Refusé | % Admis |
|------------|-------|--------|---------|
| Femmes | 1494 | 2827 | 35% |
| Hommes | 3738 | 4704 | 44% |

- ❑ Question: Les femmes ont-elles autant de chances d'être admises?

(*The Ubiquity of the Simpson's Paradox*, Selvitella, Journal of Statistical Distributions and Applications, 2017)

Le paradoxe de Simpson

❑ En séparant par département, on obtient plutôt:

| Département | Hommes applications | Hommes admis | Femmes applications | Femmes admises |
|-------------|---------------------|--------------|---------------------|----------------|
| A | 825 | 62% | 108 | 82% |
| B | 560 | 63% | 25 | 68% |
| C | 325 | 37% | 593 | 34% |
| D | 417 | 33% | 375 | 35% |
| E | 191 | 28% | 393 | 24% |
| F | 191 | 28% | 393 | 24% |

❑ Le paradoxe est dû à la distribution des genres parmi les départements.

❑ Les H/F n'appliquent pas aux mêmes endroits, et les départements n'admettent pas le même nombre de gens non plus.

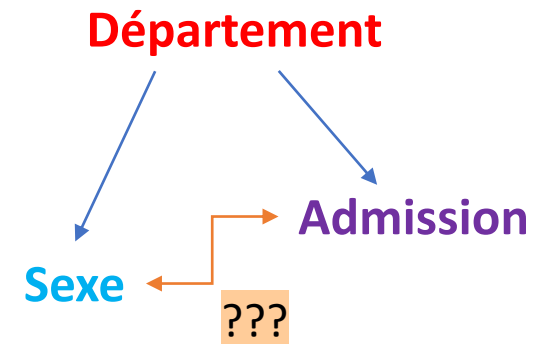
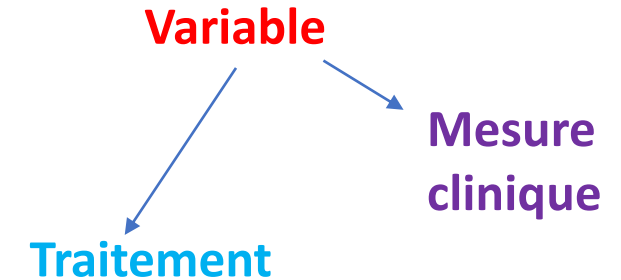
Le paradoxe de Simpson

- ❑ Dans les études en santé, ce paradoxe est fortement relié à un phénomène appelé confusion qui nous empêche d'estimer correctement des effets.

- ❑ La confusion:

- Une **variable** affecte la distribution d'un **traitement**
- La même **variable** affecte une **mesure clinique** d'intérêt
- On souhaite évaluer l'effet du **traitement** sur la **mesure clinique**.

Alors une comparaison simple de la **mesure** moyenne entre deux groupes **traitement** (ou entre les genres dans l'exemple précédent) est confondue/biaisée par cette **variable**.



Le paradoxe de Simpson

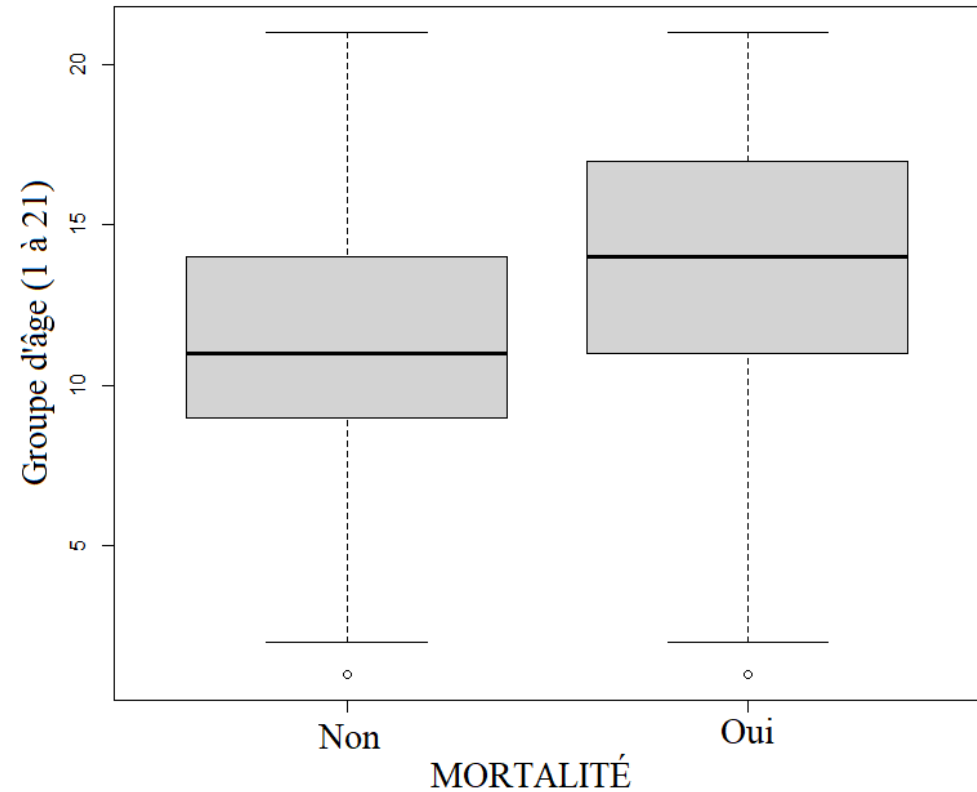
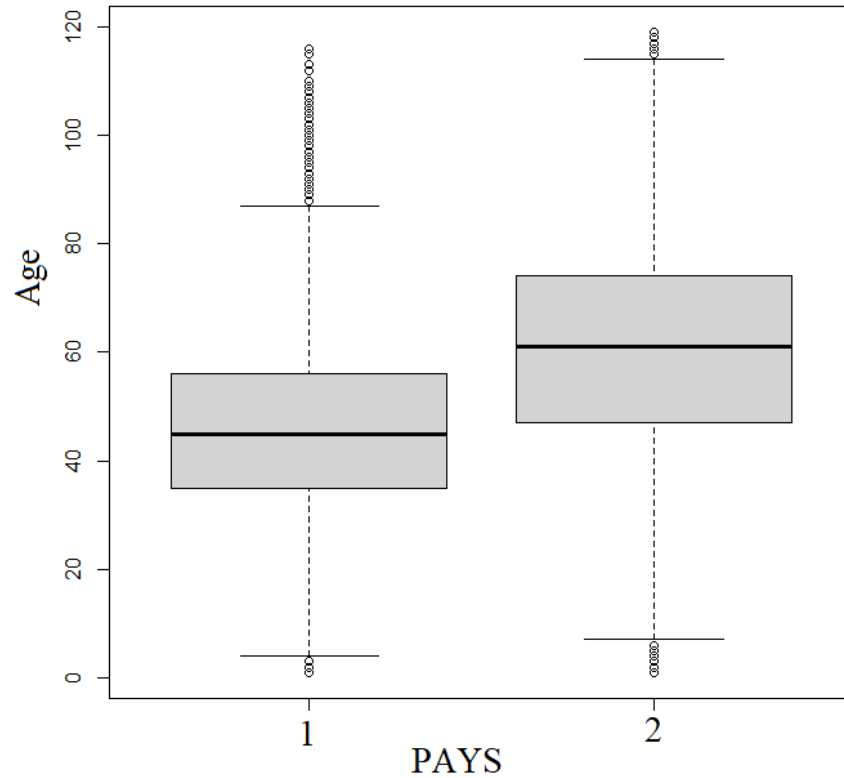
- ❑ Comme on a vu, la stratification sur le facteur de confusion peut régler le problème! (ex. stratification par département = comparaison plus juste entre les genres)
- ❑ Exemple en biostatistique: J'aimerais comparer deux populations en termes de **mortalité due à la COVID19**
- ❑ Je compare donc les taux marginaux de morts dues à la COVID19:

| Pays | Nombre de morts dues à la COVID19 | Nombre total individus | % mortalité |
|------|-----------------------------------|------------------------|-------------|
| 1 | 16 342 | 1 500 000 | 1.1% |
| 2 | 33 688 | 1 500 000 | 2.2% |

- ❑ Le pays 2 a un taux 2 fois plus élevé! Que se passe-t-il?

Le paradoxe de Simpson

❑ J'ai omis de mentionner que...



Note: Cette analyse est basée sur des nombres fictifs, les relations discutées ont été créées par simulation.

Le paradoxe de Simpson

❑ En **considérant l'âge** et en stratifiant par la tranche d'âge:

Proportion de morts dues à la COVID19 par pays

| Age \ Pays | 0-5 | 5-10 | 10-15 | 15-20 | 20-25 | 25-30 | 30-35 | 35-40 | 40-45 | 45-50 | 50-55 | 55-60 | 60-65 | 65-70 | 70-75 | 75-80 | 80-85 | 85-90 | 90-95 | 95-100 | 100+ |
|------------|-----|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------|------|
| 1 | 0.1 | 0.2 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 1.0 | 1.2 | 1.5 | 1.7 | 2.0 | 2.6 | 3.1 | 3.9 | 4.6 | 6.2 | 7.2 | 9.3 |
| 2 | 0.2 | 0.1 | 0.3 | 0.3 | 0.4 | 0.5 | 0.5 | 0.7 | 0.8 | 1.0 | 1.1 | 1.4 | 1.8 | 2.1 | 2.7 | 3.2 | 3.8 | 4.6 | 5.8 | 6.7 | 9.9 |

❑ Notez qu'il n'y a **aucune** tranche d'âge où l'on retrouve une différence de plus de 0.6 points de pourcentage entre les pays.

❑ Pour la majorité des âges, la différence est même inférieure à 0.1 point de %.

Quelques études sur la confusion ou le Paradoxe de Simpson en lien avec la COVID19...

Int J Antimicrob Agents. 2021 Apr; 57(4): 106308.

Published online 2021 Feb 17. doi: [10.1016/j.ijantimicag.2021.106308](https://doi.org/10.1016/j.ijantimicag.2021.106308)

PMCID: PMC7888988

PMID: [33609717](https://pubmed.ncbi.nlm.nih.gov/33609717/)

Perceived efficacy of hydroxychloroquine in observational studies: Results of the confounding effect of “goals of care”

Prof. Imad M. Tleyjeh, MD, MSc^{1,2,3,4,*} and Haytham Tlayjeh, MD⁵

► Author information ► Article notes ► Copyright and License information ► [Disclaimer](#)

Effects of Confounding Bias in COVID-19 and Influenza Vaccine Effectiveness Test-Negative Designs Due to Correlated Influenza and COVID-19 Vaccination Behaviors

Margaret K. Doll,^a Stacy M. Pettigrew,^a Julia Ma^b, Aman Verma,^{b,c}

^aDepartment of Population Health Sciences, Albany College of Pharmacy & Health Sciences, Albany, NY, USA; ^bPrecision Analytics, Montreal, Quebec, Canada; ^cDepartment of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada

Corresponding author: Margaret K. Doll, PhD, MPH; Albany College of Pharmacy & Health Sciences, 106 New Scotland Ave, OB210A, Albany, NY 12208 • Email: margaret.doll@acphs.edu

naturemedicine

Explore content ▼ About the journal ▼ Publish with us ▼

[nature](#) > [nature medicine](#) > [correspondence](#) > article

Correspondence | Published: 17 June 2021

Causation or confounding: why controls are critical for characterizing long COVID

Zahin Amin-Chowdhury & Shamez N. Ladhani ✉

Nature Medicine 27, 1129–1130 (2021) | [Cite this article](#)

8987 Accesses | 17 Citations | 213 Altmetric | [Metrics](#)

[J Hypertens](#). Author manuscript; available in PMC 2021 May 29.

Published in final edited form as:

[J Hypertens](#). 2021 Apr 1; 39(4): 795–805.

doi: [10.1097/HJH.0000000000002706](https://doi.org/10.1097/HJH.0000000000002706)

PMCID: PMC8164085

NIHMSID: NIHMS1695754

PMID: [33186321](https://pubmed.ncbi.nlm.nih.gov/33186321/)

Evaluating sources of bias in observational studies of angiotensin-converting enzyme inhibitor/angiotensin II receptor blocker use during COVID-19: beyond confounding

Jordana B. Cohen,^{a,b,*} Lucy D'Agostino McGowan,^{c,*} Elizabeth T. Jensen,^d Joseph Rigdon,^e and Andrew M. South^{d,f,g,h}

► Author information ► Copyright and License information ► [Disclaimer](#)

THE LANCET Respiratory Medicine

COMMENT | VOLUME 8, ISSUE 11, P1065-1066, NOVEMBER 01, 2020

Inhaled corticosteroids and COVID-19-related mortality: confounding or clarifying?

Dave Singh ✉ • David M G Halpin

Open Access • Published: September 24, 2020 • DOI: [https://doi.org/10.1016/S2213-2600\(20\)30447-1](https://doi.org/10.1016/S2213-2600(20)30447-1)

[Check for updates](#)

Simpson's paradox in Covid-19 case fatality rates: a mediation analysis of age-related causal effects

Julius von Kügelgen*, Luigi Gresele*, Bernhard Schölkopf

Abstract—We point out an instantiation of Simpson's paradox in Covid-19 case fatality rates (CFRs): comparing a large-scale study from China (17 Feb) with early reports from Italy (9 Mar), we find that CFRs are lower in Italy for every age group, but higher overall. This phenomenon is explained by a stark difference in case demographic between the two countries. Using this as a motivating example, we introduce basic concepts from mediation analysis and show how these can be used to quantify different direct and indirect effects when assuming a coarse-grained causal graph involving country, age, and case fatality. We curate an age-stratified CFR dataset with >750k cases and conduct a case study, investigating total, direct, and indirect (age-mediated) causal effects between different countries and at different points in time. This allows us to separate age-related effects from others unrelated to age and facilitates a more transparent comparison of CFRs across countries at different stages of the Covid-19 pandemic. Using longitudinal data from Italy, we discover a sign reversal of the direct causal effect in mid-March which temporally aligns with the reported collapse of the healthcare system in parts of the country. Moreover, we find that direct and indirect effects across 132 pairs of countries are only weakly correlated, suggesting that a country's policy and case demographic may be largely unrelated. We point out limitations and extensions for future work, and, finally, discuss the role of causal reasoning in the broader context of using AI to combat the Covid-19 pandemic.

Comparaison Italie-Chine

Mesure: mortalité due à la COVID19.

Résultat: Mortalité plus faible en Italie pour chaque tranche d'âge (séparément) mais, globalement, l'Italie a des taux plus élevés de mortalité que la Chine.

2

Le paradoxe de l'amitié

- ❑ Observé par Scott Feld en 1991 (études sociologiques)
- ❑ Paradoxe: Chaque personne a moins d'amis que ses amis ont d'amis, en moyenne



- ❑ Lien avec la théorie sur les réseaux

Le paradoxe de l'amitié

- ❑ Se démontre à partir de l'inégalité Cauchy-Shwarz (qui est aussi souvent utilisée en statistique dans d'autres preuves).

$$\left(\sum_{i=1}^n u_i v_i \right)^2 \leq \left(\sum_{i=1}^n u_i^2 \right) \left(\sum_{i=1}^n v_i^2 \right)$$

- ❑ Se base sur certaines hypothèses du réseau (qui sont souvent vérifiées dans le cas de réseau sociaux) mais l'inégalité n'est pas une certitude mathématique absolue.

❑ Intuition:

Les gens qui ont plusieurs amis ont plus de chance de se retrouver dans le cercle d'amis d'un membre du réseau.

Comment utiliser ce paradoxe en épidémiologie?

Le paradoxe de l'amitié

1. Pour sélectionner (plus efficacement) des gens à immuniser contre un virus

PHYSICS TODAY

HOME BROWSE▼ INFO▼ RESOURCES▼ JOBS

Home > November 2010 (Volume 63, Issue 11) > Page 15. doi:10.1063/1.3518199

Using the friendship paradox to sample a social network

When applied to random nodes in a network, the statement "Your friends have more friends than you do" has predictive power.

Mark Wilson

PDF 4 COMMENTS f t in d m e TOOLS < PREV NEXT >

Physics Today **63**, 11, 15 (2010); <https://doi.org/10.1063/1.3518199>

“ According to their simulations of computer and population networks, given a million randomly chosen nodes (computers or people), only a small fraction of random “acquaintances” of those nodes actually require immunization to arrest an unfolding computer virus or disease epidemic, compared to a large fraction needed in completely random immunization.² ”


2. Pour monitorer des gens plus efficacement dans un réseau (ex. pour comptabiliser des infections)

PLOS ONE

 OPEN ACCESS  PEER-REVIEWED

RESEARCH ARTICLE

Social Network Sensors for Early Detection of Contagious Outbreaks



Nicholas A. Christakis , James H. Fowler

Published: September 15, 2010 • <http://dx.doi.org/10.1371/journal.pone.0014008>

We therefore explore a novel, alternative strategy that does *not* require ascertainment of global network structure, namely, *monitoring the friends of randomly selected individuals*. This strategy exploits an interesting property of human social networks: on average, the friends of randomly selected people possess more links (have higher degree) and are also more central (e.g., as measured by betweenness centrality) to the network than the initial, randomly selected people who named them.^{[15]–[19]} Therefore, we expect a set of nominated friends to get infected earlier than a set of randomly chosen individuals (who represent the population as a whole).

3. Pour contenir un virus en identifiant les gens plus à risque d'être infectés.

Association Between Sampling Method and Covid-19 Test Positivity Among Undergraduate Students: Testing Friendship Paradox in Covid-19 Network of Transmission

 Sina Kianersi, Yong-Yeol Ahn,  Molly Rosenberg

doi: <https://doi.org/10.1101/2020.12.14.20248144>

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should not be used to guide clinical practice.

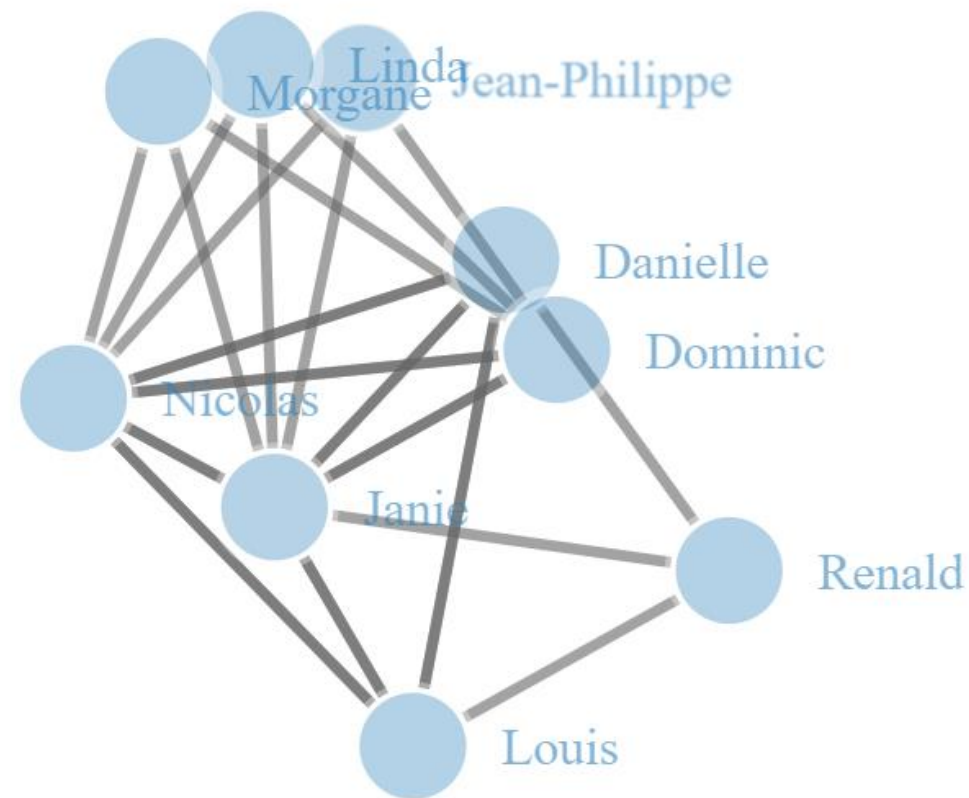
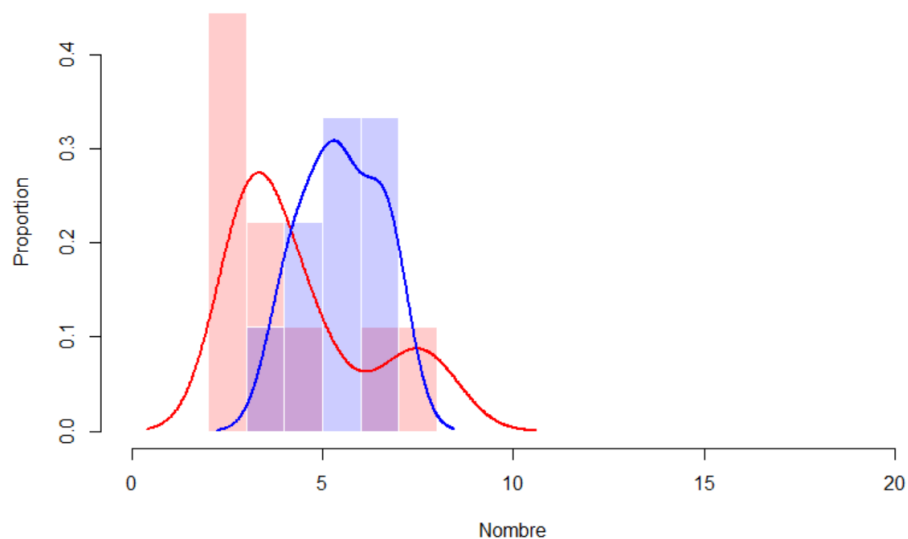


Identifying students who are more likely to get COVID-19 could improve the disease containment efforts, such as intervention programs, surveillance testing, and even vaccination strategies, and make them more efficient and effective. “Friendship Paradox” (2), a fundamental concept in network science, could potentially help to identify students with higher COVID-19 positivity risk.

We found a positive association between the sampling method and COVID-19 positivity. Nominated friends were more likely to have had COVID-19 compared to randomly sampled students.

Le paradoxe de l'amitié

- ❑ J'ai inventé un réseau inspiré par des membres de ma famille.
- ❑ Voici l'histogramme du nombre de contacts de chaque membre dans le réseau et du nombre de contacts des amis de chaque membre du réseau:



Le paradoxe de l'amitié

- ❑ J'ai obtenu les nombres de contacts moyens suivants:
 - 4.4 contacts en moyenne chez les membres du réseau et
 - 5.5 contacts en moyenne chez les amis de chaque membre du réseau.
- ❑ À partir du code R de l'atelier, vous pourrez créer votre propre réseau et vérifier le paradoxe dans votre réseau.

3

Le paradoxe de Berkson

- ❑ Le fait de se concentrer sur un sous-groupe de la population peut changer l'effet d'une variable sur une autre.
- ❑ L'effet est donc différent à l'intérieur de ce sous-groupe (parfois même inverse) par rapport à la population globale.
- ❑ Par exemple, supposons:

Je m'intéresse à l'effet d'une maladie mystérieuse (inventée!) appelée HA1P6N sur les chances d'être infecté à la COVID-19.

Prévalence de HA1P6N : 5% de la population et plusieurs seront hospitalisés.

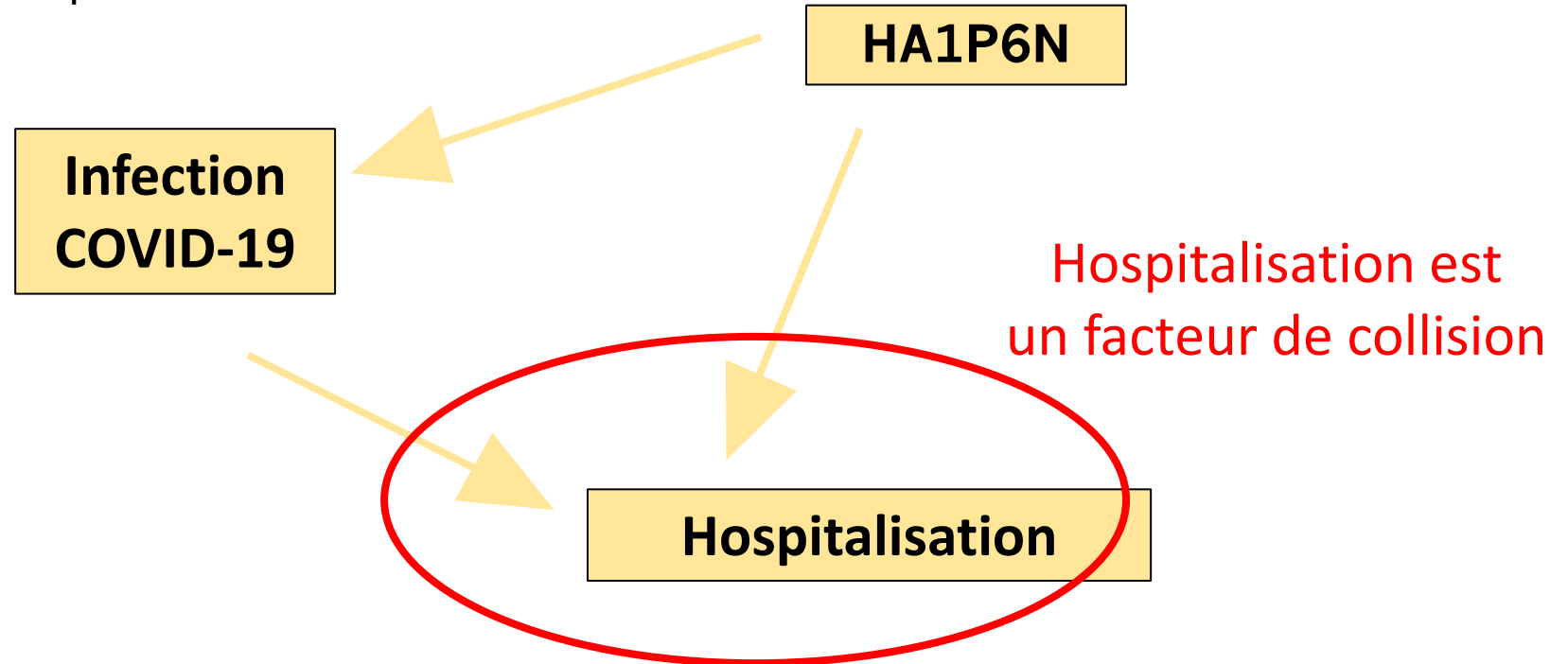
Le paradoxe de Berkson

- ❑ On a accès à un échantillon de patients qui ont été hospitalisés.
- ❑ Chez ces patients hospitalisés, on compare les taux d'infection à la COVID19 entre ceux qui ont la maladie HA1P6N et les autres.
- ❑ On trouve un ratio dans les deux groupes de 2.2, donc les patients HA1P6N ont plus de deux fois plus de chance d'avoir la COVID19:

| Patients | Total admis | Positif à COVID19 | % positif COVID19 |
|----------------|-------------|-------------------|-------------------|
| HA1P6N | 152 | 4 | 2.6% |
| Non HA1P6N | 651 | 8 | 1.2% |
| Ratio des taux | | | 2.2 > 1 |

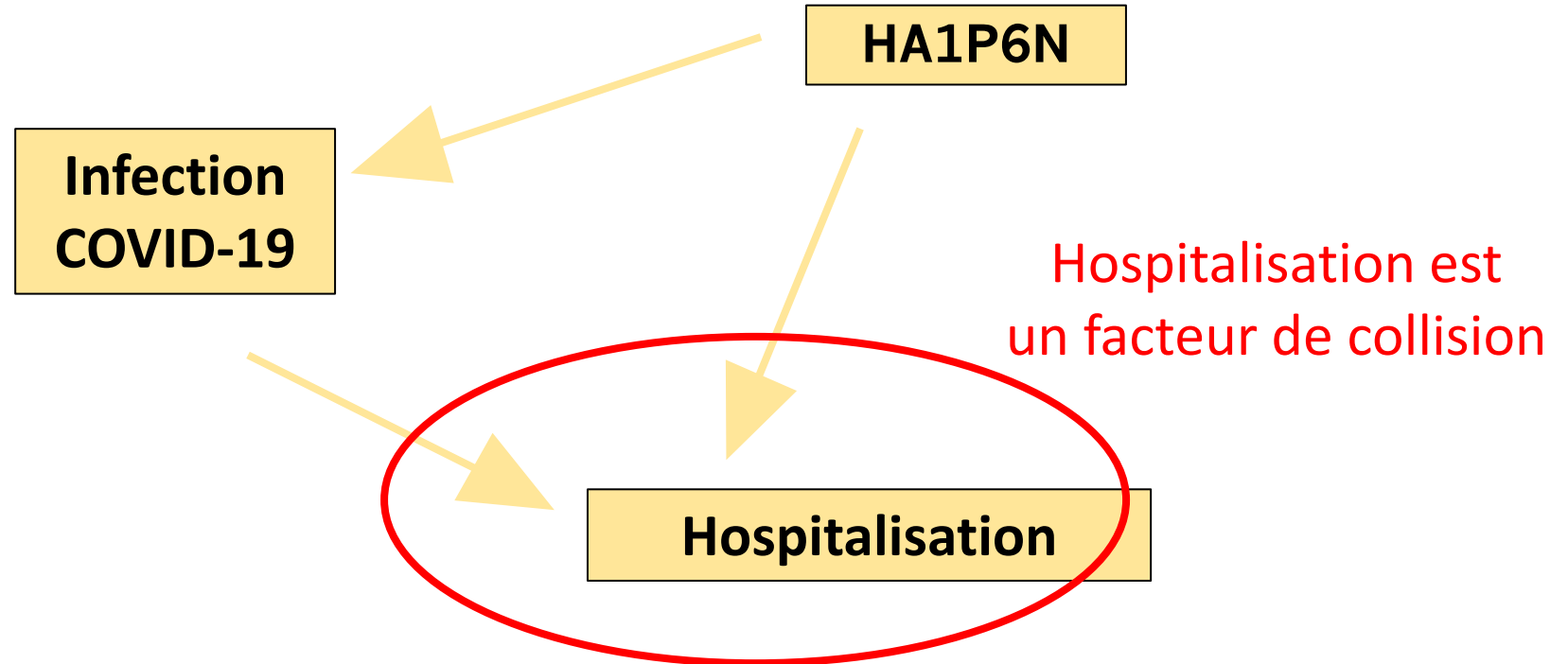
Le paradoxe de Berkson

- ❑ Notez que l'on pourrait avoir le même problème avec plusieurs maladies réelles ou autres facteurs de risque (diabète, effet de la cigarette, âge, etc.).
- ❑ On postule les relations (causales) suivantes entre la maladie HA1P6N, la COVID19 et les hospitalisations:



Le paradoxe de Berkson

- ❑ Restreindre l'analyse à cette strate de la population qui est hospitalisée crée une distortion de l'effet du HA1P6N sur la COVID 19.
- ❑ L'effet chez les hospitalisés ne peut pas être interprété comme un effet causal.



Le paradoxe de Berkson

- ❑ Si on avait accès aux données de la population entière, on verrait en fait que **l'effet est inversé**:

| Patients | Total personnes | Positif à COVID19 | % positif COVID19 |
|----------------|-----------------|-------------------|-------------------|
| HA1P6N | 5073 | 30 | 0.6% |
| Non HA1P6N | 94 917 | 659 | 0.7% |
| Ratio des taux | | | 0.9 < 1 |

- ❑ Ainsi, en utilisant les données d'hôpital, on concluerait faussement que HA1P6N est un facteur de risque pour l'infection à la COVID19.

Note: Cette analyse est basée sur des nombres fictifs, les relations discutées ont été créées par simulation.

“

There have been many surprising things written and said about the coronavirus pandemic, but perhaps none more so than the claim that smoking might protect against Covid-19 infection (bit.ly/2YLudbR). ”

”

SIGNIFICANCE

ROYAL
STATISTICAL
SOCIETY
DATA | EVIDENCE | DECISIONSASA
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of StatisticsStatistical
Society of
AustraliaNotebook | [Free Access](#)

The spectre of Berkson's paradox: Collider bias in Covid-19 research

Annie Herbert, Gareth Griffith, Gibran Hemani, Luisa Zuccolo

First published: 29 July 2020 | <https://doi.org/10.1111/1740-9713.01413> | Citations: 6

Dans l'étude discutée, on restreignait aussi l'analyse aux gens en hôpital!

❑ Aussi appelé biais dû à la stratification sur un facteur de collision

Ici, on parle de facteurs de collision comme l'hospitalisation, ou les gens qui se font tester (donc qui soupçonnent déjà la maladie), ou des gens volontaires dans une étude.

nature communications

Explore content ▾

About the journal ▾

Publish with us ▾

[nature](#) > [nature communications](#) > [articles](#) > article
Article | [Open Access](#) | [Published: 12 November 2020](#)

Collider bias undermines our understanding of COVID-19 disease risk and severity

[Gareth J. Griffith](#), [Tim T. Morris](#), [Matthew J. Tudball](#), [Annie Herbert](#), [Giulia Mancano](#), [Lindsey Pike](#),

[Gemma C. Sharp](#), [Jonathan Sterne](#), [Tom M. Palmer](#), [George Davey Smith](#), [Kate Tilling](#), [Luisa Zuccolo](#), [Neil](#)
[M. Davies](#) & [Gibran Hemani](#) ✉

Conclusion

Nous avons vu 3 paradoxes surprenants:

Simpson:

- La distribution d'une troisième variable (département, âge, etc.) qui affecte/cause les deux variables qui nous intéressent peut biaiser notre inférence!
- Stratifier sur cette variable confondante peut régler le problème.

Amitié:

- Les personnes dans un réseau ont moins d'amis que leurs amis.
- On peut utiliser ce paradoxe à bon escient!

Berkson:

- Stratifier sur un sous-groupe d'une variable de collision (ex. les hospitalisations) peut mener à une distortion de l'effet d'une autre variable.

Conclusion

- ❑ Chaque paradoxe peut s'expliquer par des phénomènes connus en (bio)statistique ou en épidémiologie.
- ❑ La connaissance de ces paradoxes permet de mieux cerner des problèmes potentiels dans les analyses statistiques de données réelles.
- ❑ **En deuxième partie, nous reproduirons les 3 analyses discutées et verrons ces paradoxes en action!**