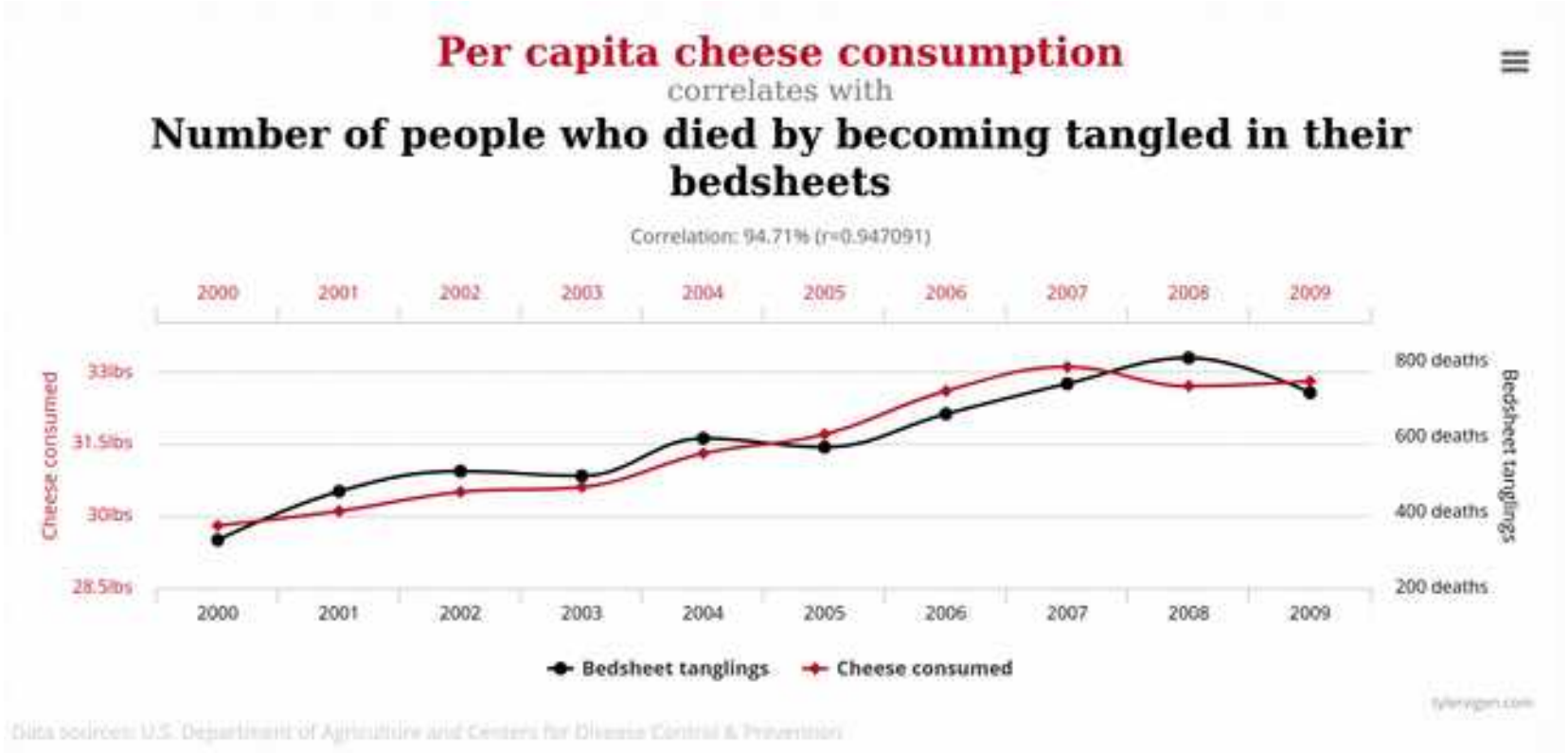
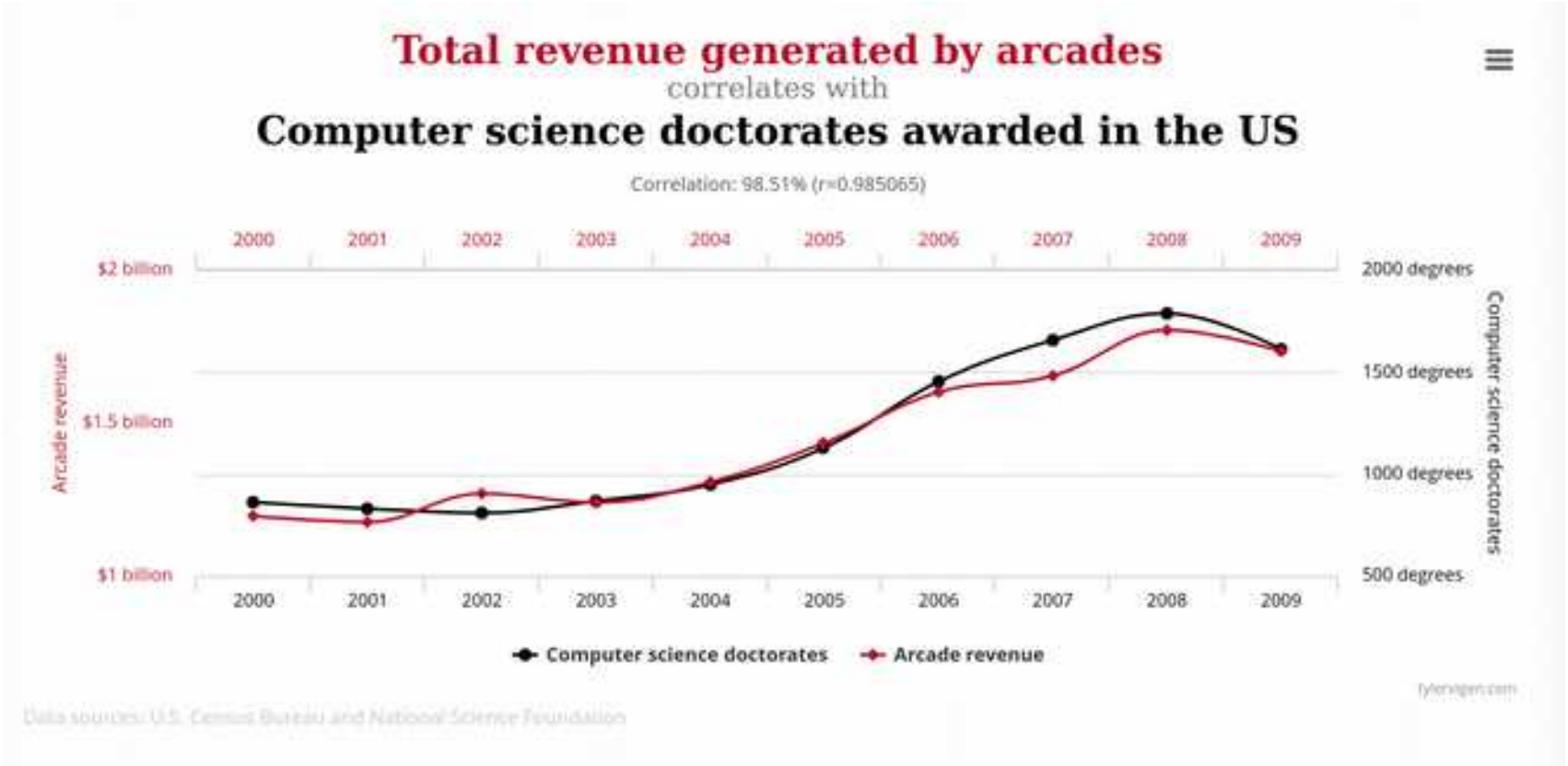
PARADOXES STATISTIQUES

Ou comment prouver n’importe quoi avec une certitude absolue


Per capita cheese consumption
correlates with
Number of people who died by becoming tangled in their bedsheets
Correlation: 94.71% (r=0.947091)
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Cheese consumed
33lbs
31.5lbs
30lbs
28.5lbs
800 deaths
600 deaths
400 deaths
200 deaths
Bedsheet tanglings
Bedsheet tanglings
Cheese consumed


Total revenue generated by arcades
correlates with
Computer science doctorates awarded in the US
Correlation: 98.51% (r=0.985065)
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
$2 billion
$1.5 billion
$1 billion
Arcade revenue
2000 degrees
1500 degrees
1000 degrees
500 degrees
Computer science doctorates
Computer science doctorates
Arcade revenue
tylervigen.com
Data sources: U.S. Census Bureau and National Science Foundation

Data dredging (“triturage de données”)

Je formule l’hypothèse suivante: “la probabilité qu’une pièce retombe sur face est de 60%”.

Comment la prouver? Il faut lancer des pièces et voir si on obtient un résultat proche de 60%

Ça tombe bien! J’ai justement fait l’expérience hier soir, j’ai lancé 20 pièces et j’ai obtenu 12 “face” et 8 “pile, soit une probabilité de 60%!

Data dredging (“triturage de données”)

Expérience n°2: je demande à 10000 personnes de répondre à un questionnaire. J’étudie les résultats, et je peux publier ce super papier:

Les personnes de 37 ans ont une probabilité beaucoup plus élevée que les autres d’avoir un grain de beauté sur la fesse gauche

Etude réalisée sur 10000 personnes formant un échantillon représentatif de la population, p = 0.001

Indice: mon questionnaire avait 1000 questions...

Data dredging (“triturage de données”)

Problème: on a élaboré les hypothèses **à posteriori**, c’est à dire en se basant sur les données, et on a confirmé l’hypothèse en utilisant **les mêmes données**!!

La démarche paraît évidemment problématique pour le lancer de pièces, mais pour l’étude multicritères c’est déjà beaucoup moins clair…

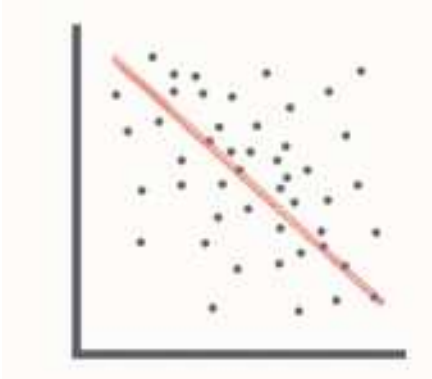
$$\text{Proba(trouver une fausse corrélation)} \xrightarrow[\text{nb paramètres} \to \infty]{} 1$$

Solution: -Soumettre l’hypothèse **avant** de faire l’étude

-Diviser l’échantillon en 2: 1 pour formuler une hypothèse, l’autre pour la tester

Paradoxe de Berkson

On demande à des personnes de noter la qualité des burgers et des frites du dernier fast-food où ils sont allés. On obtient le résultat suivant:



Il semble donc que la qualité des frites soit négativement corrélée à celle des burgers…

Paradoxe de Berkson

Problème: Si un fast-food a de mauvaises frites ET de mauvais burgers, les gens n'y vont pas

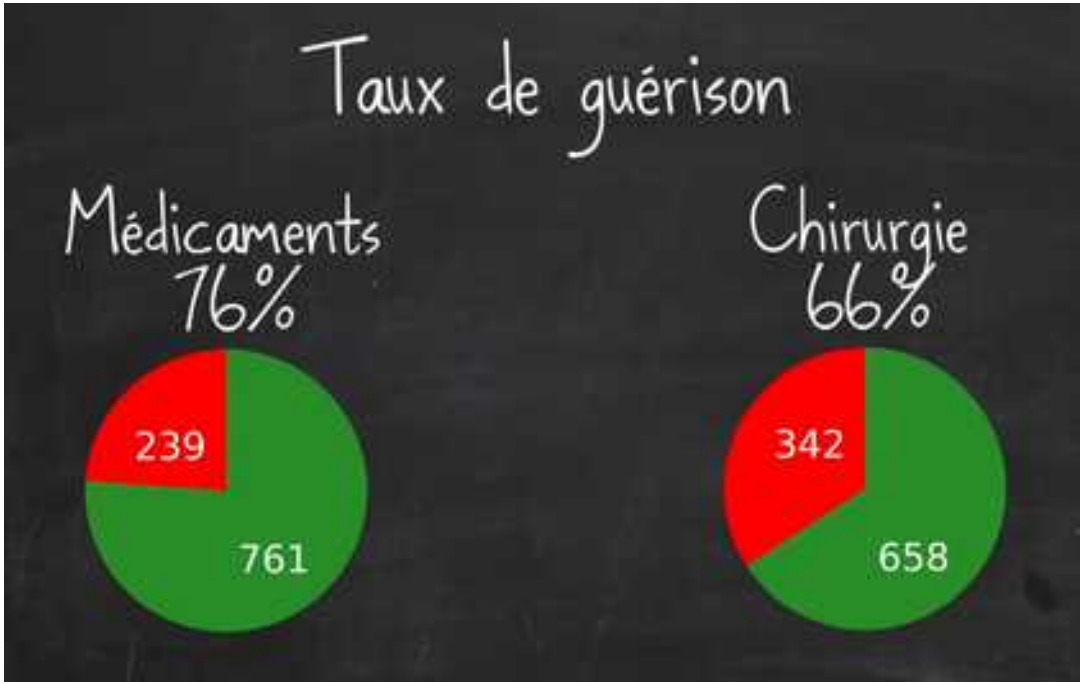
 Cela revient à retirer ces restaurants de l'étude, et donc l'échantillon devient biaisé

Cause: Si on étudie le lien entre X et Y, et qu'on exclut les individus pour lesquels X et Y sont simultanément trop bas, on crée artificiellement une corrélation négative

Solution: Il faut s'assurer que les critères étudiés sont **indépendants** des critères sur lesquels on a sélectionné les participants à l'étude.

Paradoxe de Simpson

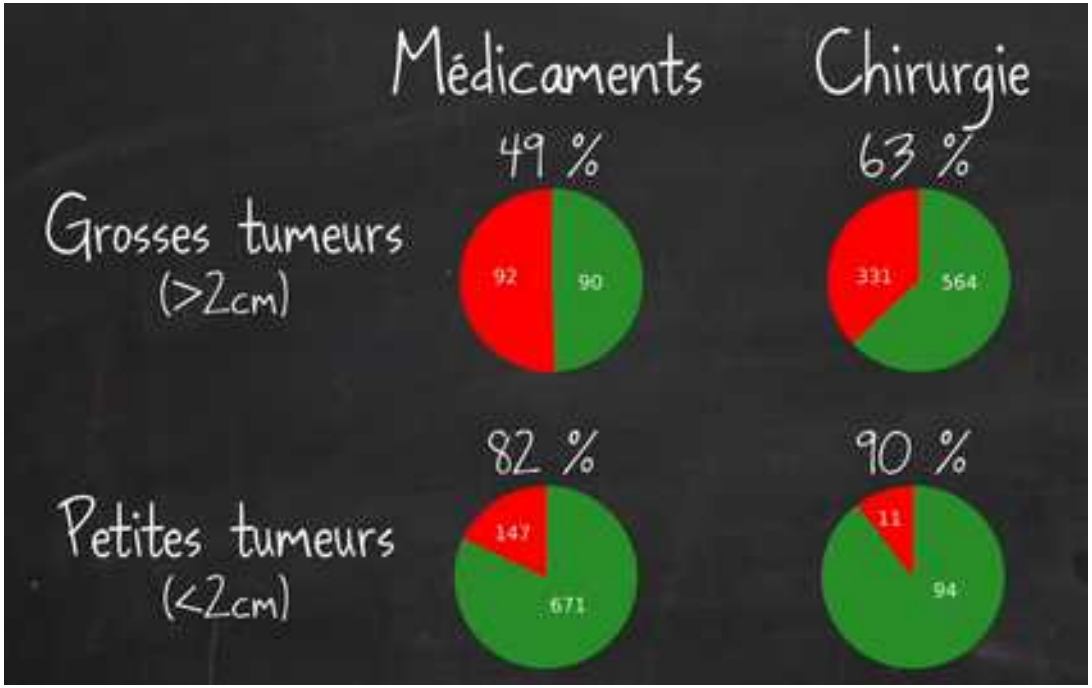
Votre médecin vous diagnostique un cancer. Il existe deux traitements: la chirurgie, ou la chimio. Vous vous demandez lequel est le plus efficace, voici les chiffres:



Source:
ScienceEtonnante

Paradoxe de Simpson

Votre oncologue vous indique qu'il faut tout de même prendre en compte la taille de la tumeur. Vous cherchez donc les statistique correspondantes:

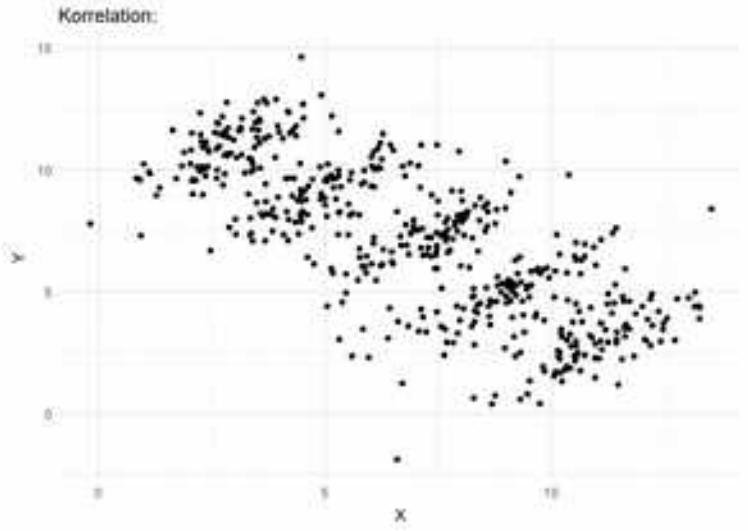


Source:
ScienceEtonnante

Paradoxe de Simpson

Problème: La chirurgie est meilleure quelle que soit la taille de la tumeur, mais elle est moins efficace dans l’ensemble (?!)

Explication: La chirurgie est beaucoup plus utilisée sur les grosses tumeurs, qui sont plus graves, alors que la chimio est plus souvent utilisée sur les petites tumeurs.



Paradoxe de Simpson

Causes: 1) **Facteurs de confusion** - un paramètre qui joue sur plusieurs des paramètres qu'on veut étudier

2) **Distribution non homogène** - pas le même nombre de points de données pour chaque variable de part et d'autre du facteur de confusion

Solution: -Demander à un expert d'identifier les possibles facteurs de confusion

-Faire des études **prospectives** et non pas **rétrospectives**

Prospectif: On veut tester une hypothèse, on fait une expérience en veillant à la **représentativité** des échantillons
Rétrospectif: On étudie des données existantes et inhomogènes

Conclusion

-Un très grand nombre d’études statistiques qu’on voit tous les jours tombent dans un ou plusieurs de ces biais;

-Il est très facile pour les créateurs de l’étude de cacher ces biais, et très difficile pour l’observateur de s’en rendre compte;

-En tant que scientifique, il existe des bonnes pratiques pour ne pas tomber dans ces pièges involontairement;

-En tant qu’observateur, connaître ces biais permet partiellement de les détecter, mais il est indispensable d’avoir un expert avec soi pour trouver les facteurs de confusion.