Bonnes pratiques en statistiques

Il y a trois types de mensonges : les mensonges, les sacrés mensonges et les statistiques (Marc Twain)

Thomas LALOË VERDELHAN

08 décembre 2023

Université Côte d'Azur

Rendons à César ...

- Chaine Youtube « science étonnante » :

  `https://www.youtube.com/c/ScienceEtonnante`

- Chaine Youtube « la biologie fait des vidéos » :

  `https://www.youtube.com/c/LaBiologiefaitdesvidÃl'os`

- Blog « Data Visualisation » :

  `http://data.visualisation.free.fr/`

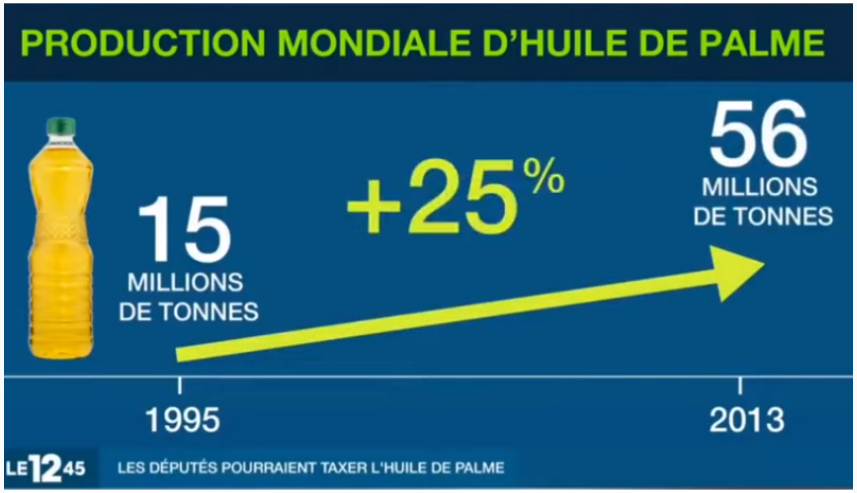
- Collectif « Le Cortecs » :

  `https://cortecs.org/`

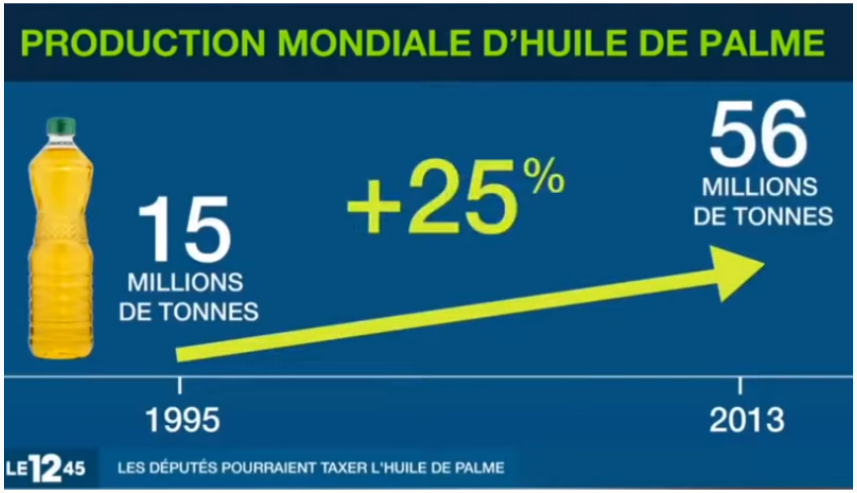
Sommaire

Pourcentages et probabilités, non à la maltraitance !

Pourcentage d'augmentation

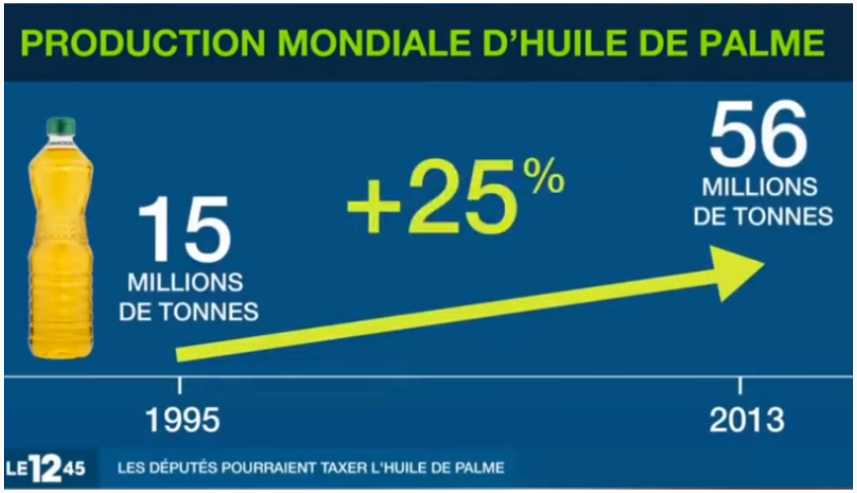


Pourcentage d'augmentation



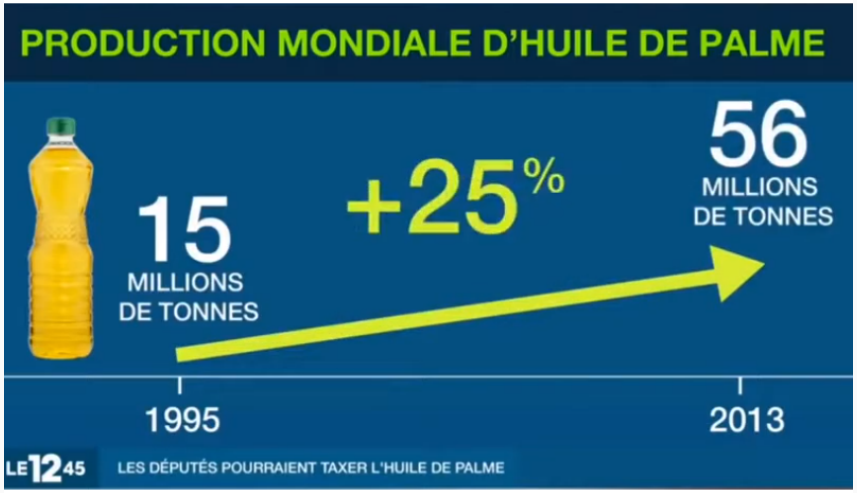
- Augmentation : $\frac{56}{15} = 3.73$, soit $100 * (3.73 - 1) = +273\%$

Pourcentage d'augmentation



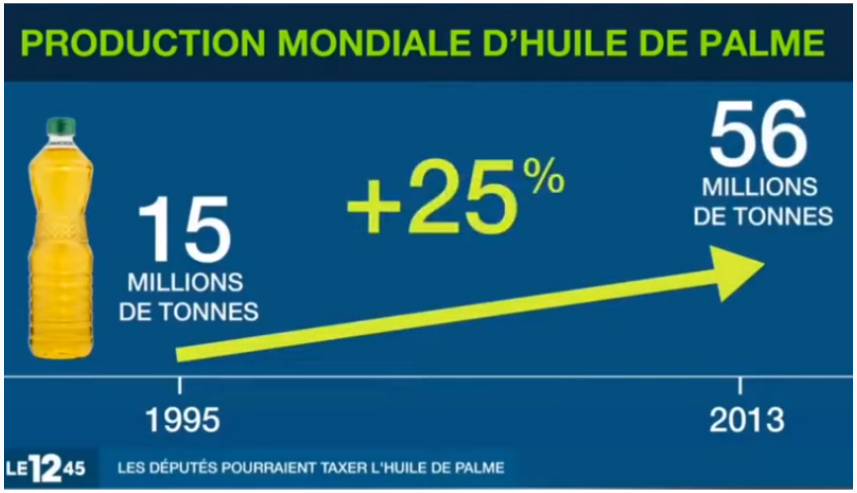
- Augmentation : $\frac{56}{15} = 3.73$, soit $100 * (3.73 - 1) = +273\%$
- Diminution : $\frac{15}{56} = 0.27 \approx 0.25$, soit $100 * (0.27 - 1) = -73\%$

Pourcentage d'augmentation



- Augmentation : $\frac{56}{15} = 3.73$, soit $100 * (3.73 - 1) = +273\%$
- Diminution : $\frac{15}{56} = 0.27 \approx 0.25$, soit $100 * (0.27 - 1) = -73\%$
- Première erreur : $\frac{\mathit{initial}}{\mathit{final}} \neq \frac{\mathit{final}}{\mathit{initial}}$

Pourcentage d'augmentation



- Augmentation : $\frac{56}{15} = 3.73$, soit $100 * (3.73 - 1) = +273\%$
- Diminution : $\frac{15}{56} = 0.27 \approx 0.25$, soit $100 * (0.27 - 1) = -73\%$
- Première erreur : $\frac{initial}{final} \neq \frac{final}{initial}$
- Deuxième erreur : le pourcentage d'évolution est

$$100 * (\frac{final}{initial} - 1)$$

Addition de pourcentages

Addition de pourcentages :

« Le département a augmenté les impôts de 30%, la région de 58%, c'est la double peine, pof 88% d'augmentation »(V. Pécresse, ministre de l'enseignement supérieur et de la recherche)

Addition de pourcentages

Addition de pourcentages :

« Le département a augmenté les impôts de 30%, la région de 58%, c'est la double peine, pof 88% d'augmentation »(V. Pécresse, ministre de l'enseignement supérieur et de la recherche)

Comment faire ?

Une seule solution, revenir au données brutes :

	Initial	Final
Région	100 €	130 € (+30%)
Département	50€	79€ (+58%)
Total	150€	209€ (+39%)

	Initial	Final
Région	50€	65 € (+30%)
Département	100€	158€ (+58%)
Total	150€	223€ (+48%)

Probabilités conditionnelles

« 30% des patients en réanimation ont moins de 60 ans. C'est-à-dire que si vous mettez un âge à 60 ans, vous vous dites : ce n'est pas grave, il y a 30% de gens qui ont moins de 60 ans (qui ne seront pas confinés) qui seront mis en danger et qui pourront continuer à aller en réanimation. »O. Véran, Ministre de la santé.

Probabilités conditionnelles

« 30% des patients en réanimation ont moins de 60 ans. C'est-à-dire que si vous mettez un âge à 60 ans, vous vous dites : ce n'est pas grave, il y a 30% de gens qui ont moins de 60 ans (qui ne seront pas confinés) qui seront mis en danger et qui pourront continuer à aller en réanimation. »O. Véran, Ministre de la santé.

$\Rightarrow$ Nous avons donc une confusion entre $\mathbb{P}(\mathit{Hospitalisation}|\mathit{Age})$ et $\mathbb{P}(\mathit{Age}|\mathit{Hospitalisation})$.

Probabilités conditionnelles

« 30% des patients en réanimation ont moins de 60 ans. C'est-à-dire que si vous mettez un âge à 60 ans, vous vous dites : ce n'est pas grave, il y a 30% de gens qui ont moins de 60 ans (qui ne seront pas confinés) qui seront mis en danger et qui pourront continuer à aller en réanimation. »O. Véran, Ministre de la santé.

$\Rightarrow$ Nous avons donc une confusion entre $\mathbb{P}(\mathit{Hospitalisation}|\mathit{Age})$ et $\mathbb{P}(\mathit{Age}|\mathit{Hospitalisation})$.

$\rightarrow$ McDonalds : « Plus de 70% de nos managers et directeurs adjoints ont débuté comme équipiers ».

Probabilités conditionnelles

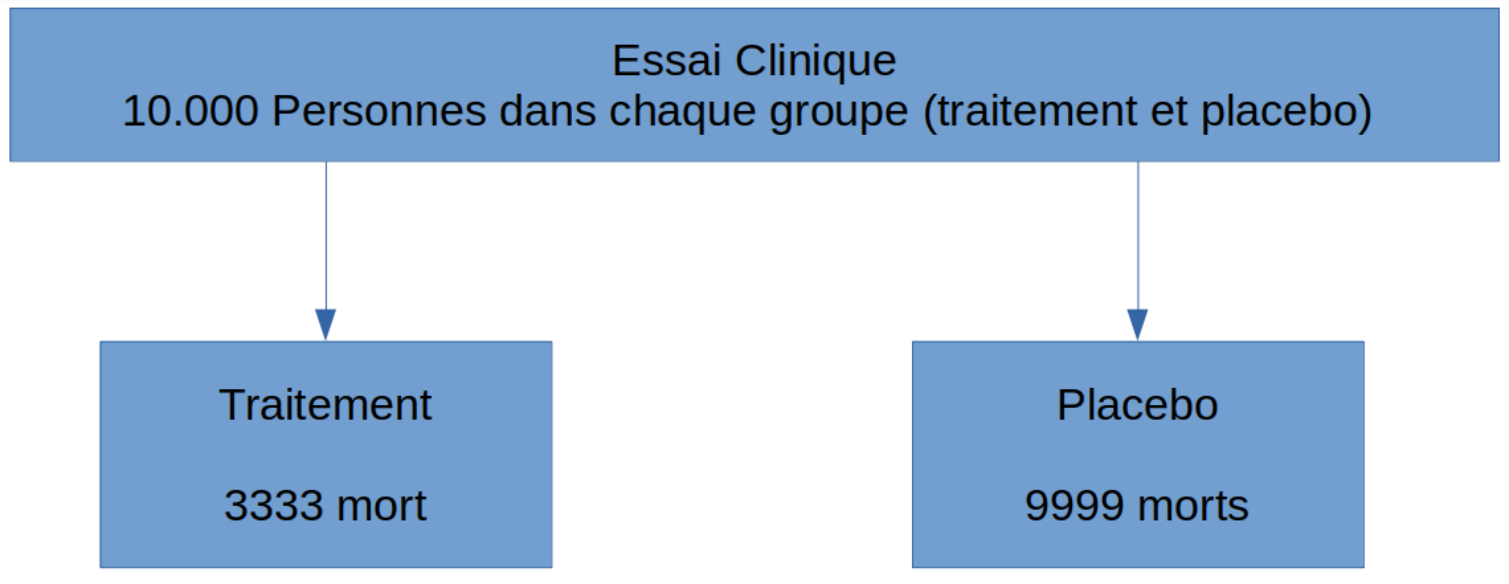
« 30% des patients en réanimation ont moins de 60 ans. C'est-à-dire que si vous mettez un âge à 60 ans, vous vous dites : ce n'est pas grave, il y a 30% de gens qui ont moins de 60 ans (qui ne seront pas confinés) qui seront mis en danger et qui pourront continuer à aller en réanimation. »O. Véran, Ministre de la santé.

$\Rightarrow$ Nous avons donc une confusion entre $\mathbb{P}(Hospitalisation|Age)$ et $\mathbb{P}(Age|Hospitalisation)$.

$\rightarrow$ McDonalds : « Plus de 70% de nos managers et directeurs adjoints ont débuté comme équipiers ».

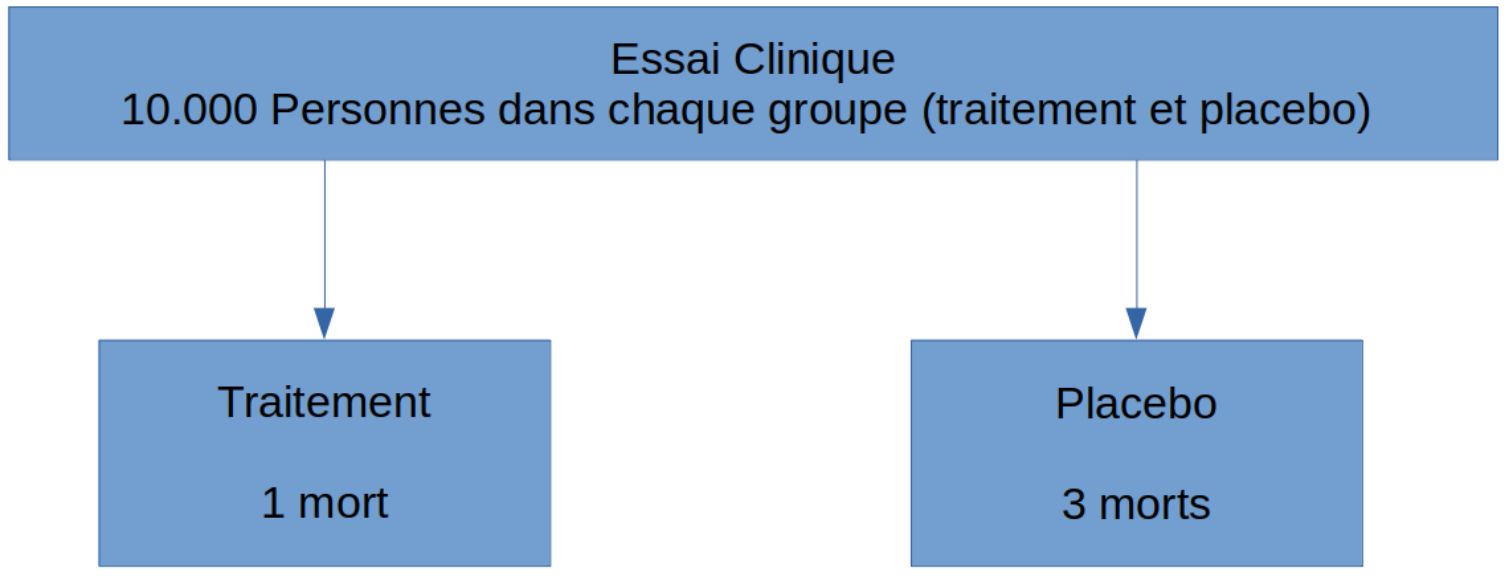
$\Rightarrow$ Est-ce la probabilité vraiment intéressante ?

Risque Absolu - Risque relatif



- Réduction du risque relatif : $(1 - \frac{3333}{9999}) * 100 = 66.7\%$
  $\Rightarrow$ 3 fois moins de chances de mourir avec le traitement
- Réduction du risque absolu : $\frac{9999-3333}{10000} * 100 = 66.7\%$
  $\Rightarrow$ Le traitement permet de sauver 2 personnes sur 3

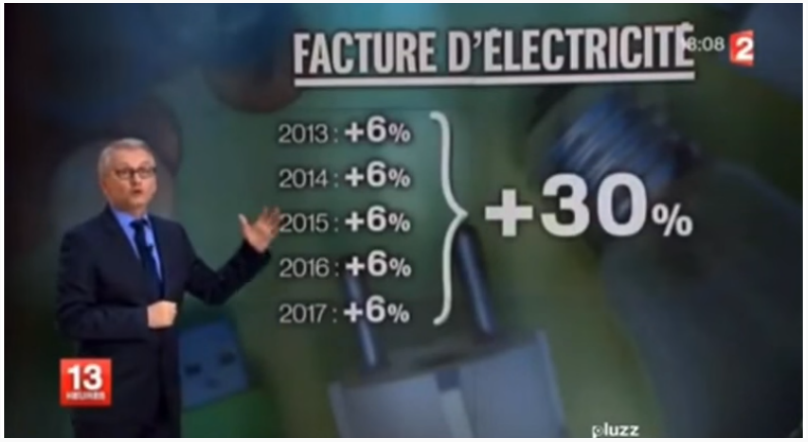
Risque Absolu - Risque relatif



- Réduction du risque relatif : $(1 - \frac{1}{3}) * 100 = 66.7\%$
  $\Rightarrow$ 3 fois moins de chances de mourir avec le traitement
- Réduction du risque absolu : $\frac{3-1}{10000} * 100 = 0.02\%$
  $\Rightarrow$ Le traitement permet de sauver 2 personnes sur 10000 (0.02%)

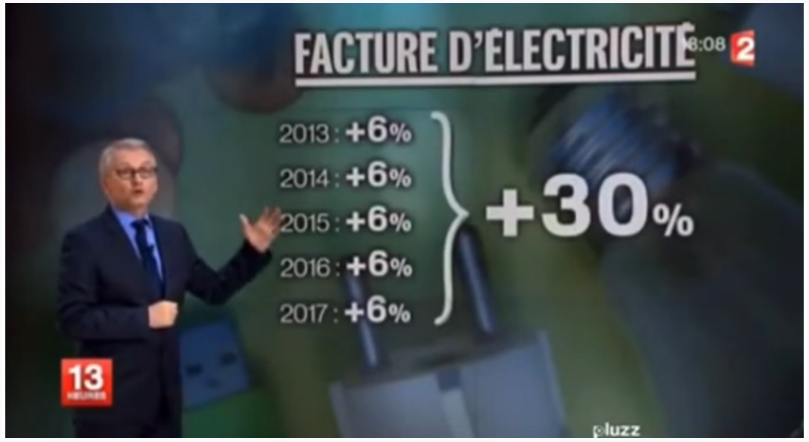
Addition de pourcentages

Augmentation cumulée :



Addition de pourcentages

Augmentation cumulée :



Le bon calcul :

	Départ	2013	2014	2015	2016	2017
Prix	100	106	112.36	119.1	126.2	133.82
Augmentation cumulée		6.00%	12.36%	19.10%	26.20%	33.82%

Bilan

- Additions, cumuls, mauvais usages, erreurs probabilistes ... Les dangers sont partout ;
- Toujours se méfier quand on entend parler de pourcentages ;
- Ne pas hésiter à revenir aux données de départ.

Accident nucléaire : une certitude statistique !

Une tribune dans Libération

TRIBUNE

Accident nucléaire : une certitude statistique

par Bernard LAPONCHE, hysicien nucléaire, expert en politiques de lénergie et Benjamin Dessus, Ingénieur et économiste, président de Global Chance

publié le 3 juin 2011 à 0h00

« Sur la base du constat des accidents majeurs survenus ces trente dernières années, la probabilité d'occurrence d'un accident majeur sur ces parcs serait donc de 50% pour la France et de plus de 100% pour l'Union européenne. Autrement dit, on serait statistiquement sûr de connaître un accident majeur dans l'Union européenne au cours de la vie du parc actuel et il y aurait une probabilité de 50% de le voir se produire en France. »

Le calcul fait

- Estimation sur les 30 dernières années : 14000 réacteurs-ans observés, quatre accidents majeurs $\rightarrow$ une probabilité d'accident pour un réacteur sur une période d'un an : $p = 0.0003$;

- 143 réacteurs en activité en Europe, donc sur une période de 30 ans on aura $143 * 30 = 4290$ réacteurs-ans ;

- Probabilité d'un accident majeur en Europe sur 30 ans : $4290 * 0.0003 = 1.29$. Il y aurait donc 129% de chance qu'un réacteur explose en Europe pendant les 30 prochaines années.

Avec un exemple plus simple : tir à pile ou face

- On tire @ pile ou face : on a une probabilité $p = 0.5$ d'avoir pile

- On s'intéresse à la probabilité d'avoir au moins un pile sur 3 lancés de pièces ;

- $\mathbb{P}(\#Piles \geq 1) = 3 * 0.5 = 1.5$. Il y aurait donc 150% de chances de faire au moins un pile sur trois lancés.

Un peu de probabilités

Loi Bernouilli :

X suit une loi de Bernoulli $B(p)$ de paramètre p si :

- X prend les valeurs 0 ou 1 ;
- $\mathbb{P}(X = 1) = 1 - \mathbb{P}(X = 0) = p$;
- $\mathbb{E}(X) = p$

Un peu de probabilités

Loi Bernouilli :

X suit une loi de Bernoulli $B(p)$ de paramètre p si :

- X prend les valeurs 0 ou 1 ;
- $\mathbb{P}(X = 1) = 1 - \mathbb{P}(X = 0) = p$;
- $\mathbb{E}(X) = p$

Loi Binomiale :

- $X_1, X_2, \ldots, X_n$ sont n variables aléatoires indépendantes de même loi $B(p)$;
- $Y_n = \sum_{i=1}^{n} X_i$ suit une loi Binomiale $B(n, p)$;
- $\mathbb{P}(Y_n = k) = \binom{n}{k} * p^k * (1 - p)^{(n-k)}$;
- $\mathbb{P}(Y_n \geq 1) = 1 - \mathbb{P}(Y_n = 0) = 1 - (1 - p)^n$;
- $\mathbb{E}(Y_n) = n * p$.

Modélisation des accidents nucléaires

Soit X une variable aléatoire représentant l'évènement « un réacteur a un accident pendant en une année » :

- On peut supposer que X est distribuée selon une loi Bernoulli $B(p)$;
- p est inconnue, il faut l'estimer ;
- On s'intéresse au nombre d'accidents sur 30 ans pour 143 réacteurs ;
- Idée : modéliser le nombre d'accidents par une variable de loi Bernoulli Y_n et calculer $\mathbb{P}(Y_n \geq 1)$.

Les hypothèses nécessaires

- Les X_i sont indépendants ;
- p est constante dans le temps ;
- p est correctement estimée.

Calcul final

- Sur 30 ans on compte 14000 $(30 * 450)$ réacteurs-ans dans le monde, et 4 accidents majeurs ;
- p est la probabilité d'avoir un accident sur un réacteur pendant un an $\Rightarrow$ estimée par la fréquence, soit $4/14000 = 0.0003$.
- Sur une période de 30 ans à venir en Europe on compte $30 * 143 = 4290$ réacteurs-ans. ;
- Y_{4290} suit une loi Binomiale $B(4290, 0.0003)$;
- $\mathbb{P}(Y_{4290} \geq 1) = 1 - (1 - 0.0003)^{4290} = 0.72$;
- $\mathbb{E}(Y_{4290}) = 4290 * 0.0003 = 1.29$.

Il y aurait donc 72% de chances qu'au moins un réacteur explose en Europe pendant les 30 prochaines années.

Une autre approche : intervalles de confiance

- Plutôt que de se contenter d'une estimation ponctuelle de p on pourrait considérer des intervalles de confiance ;
- X variable aléatoire Bernouilli de paramètre p, $Y_n = \sum_{i=1}^{n} X_i$ de loi Binomiale $B(n, p)$;
- On va avoir besoin de l'espérance et la variance de X (et par déduction de Y_n) pour calculer les intervalles ;
- $\mathbb{E}(X) = n * p$, $\sigma^2(X) = p * (1 - p)$:
  - Si p est faible (par exemple 0.1) ou forte (par exemple 0.9), on aura beaucoup de 0 (p faible) et 1 (p forte), donc peu de variabilité $\Rightarrow \sigma^2(X) = 0.9 * 0.1 = 0.009$) ;
  - Si $p = 0.5$, on aura en moyenne autant de 0 ou 1, et donc une variabilité maximale : $\Rightarrow \sigma^2(X) = 0.5 * 0.5 = 0.25$) ;
  - Plus il y a de variabilité et plus l'estimation de p sera difficile. Les intervalles seront donc plus grands.

Une autre approche : intervalles de confiance

Intervalle de fluctuation :

- Considérant un modèle connu, et en supposant qu'il est vrai on s'intéresse à un intervalle dans lequel les données ont une grande probabilité de tomber ;
- Notons $\mathfrak{p}$ le paramètre d'intérêt, $\hat{\mathfrak{p}}$ son estimateur, et α le niveau de confiance ;
- Un intervalle de fluctuation $IF_{n,\alpha}$ est de la forme

$$[\mathfrak{p} - q_\alpha \sigma(\mathfrak{p}); \mathfrak{p} + q_\alpha \sigma(\mathfrak{p})]$$

- On obtient $\mathbb{P}(\hat{\mathfrak{p}} \in IF_{n,\alpha}) \geq \alpha$;
- En supposant le modèle théorique vrai, et pour chaque jeu de n observations de notre variable aléatoire, on a une probabilité d'au moins α que l'estimateur $\hat{\mathfrak{p}}$ soit dans cet intervalle.

Une autre approche : intervalles de confiance

Intervalle de Confiance :

- Considérant un modèle connu, et en supposant qu'il est vrai on s'intéresse à un intervalle dans lequel les données ont une grande probabilité de tomber ;
- Notons $\mathfrak{p}$ le paramètre d'intérêt, $\hat{\mathfrak{p}}$ son estimateur, et α le niveau de confiance ;
- Un intervalle de confiance $IC_{n,\alpha}$ est de la forme

$$[\hat{\mathfrak{p}} - q_\alpha \sigma(\mathfrak{p}); \hat{\mathfrak{p}} + q_\alpha \sigma(\mathfrak{p})]$$

- On obtient $\mathbb{P}(\mathfrak{p} \in IC_{n,\alpha}) \geq \alpha$;
- En supposant le modèle théorique vrai, et pour chaque jeu de n observations de notre variable aléatoire, on a une probabilité d'au moins α que le vrai $\mathfrak{p}$ soit dans cet intervalle.

Une autre approche : intervalles de confiance

- On sait que si les tirages de X sont indépendants et que n est assez grand Y_n se comporte comme une loi normale $\Rightarrow$ Facile de calculer les intervalles ;
- On peut estimer p par la fréquence $\hat{p} = Y_n/n$;
- Intervalle de fluctuation de niveau 95% :

$$IF_{0.95} = \left[p - 2\sqrt{\frac{p(1-p)}{n}}; p + 2\sqrt{\frac{p(1-p)}{n}} \right]$$

- Intervalle de confiance de niveau 95% :

$$IC_{0.95} = \left[\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

Une autre approche : intervalles de confiance

- Retour sur l'exemple du risque nucléaire ;
- f est très faible, l'intervalle pour p va être petit ;
- Au niveau de confiance de 95% on obtient que $p \in [0.00001; 0.0006]$;
- Ce qui donne un 95% de chances que la probabilité d'avoir un accident soit dans l'intervalle $[4\%; 92\%]$. Pourquoi si grand ? ;
- $\mathbb{P}(Y_n \geq 1) = 1 - ((1-p)^n)$. Comme p est petit, $1-p$ est proche de 1, donc $(1-p)^n$ décroît lentement. Avec $n = 10^6$ réacteurs-ans on aurait l'intervalle $[99,99\%; 100\%]$
- C'est très souvent utile de présenter les intervalles de confiance (sondages par exemple).

Discussion des hypothèses

- Les accidents sont-ils indépendants ? Très certainement non. Dans les 4 accidents majeurs recensés il y en a 3 à Fukushima (une centrale peut avoir plusieurs réacteurs).

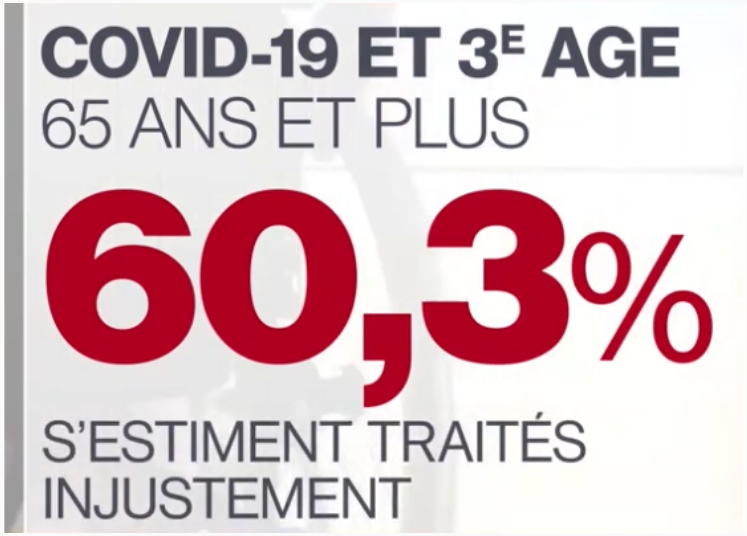
- p est il constant dans le temps ? Très certainement non, chaque accident entraîne a priori un changement des normes de sécurité ;

- p est-il bien estimé ? Difficile à dire, mais l'estimation faite fait l'hypothèse qu'il est le même dans tous les pays du monde (il est estimé sur les 14000 réacteurs-ans répartis sur le globe).

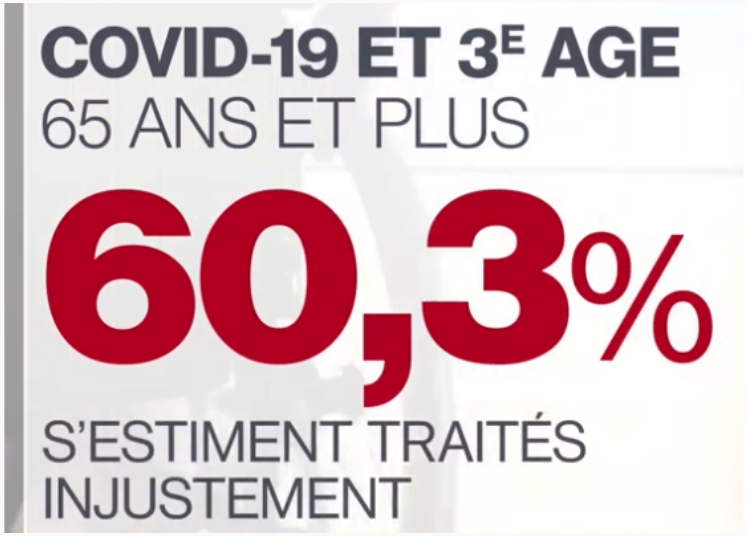
- Et si on changeait les intervalles de temps considérés ?

Biais de sélection

Enquête Covid



Enquête Covid



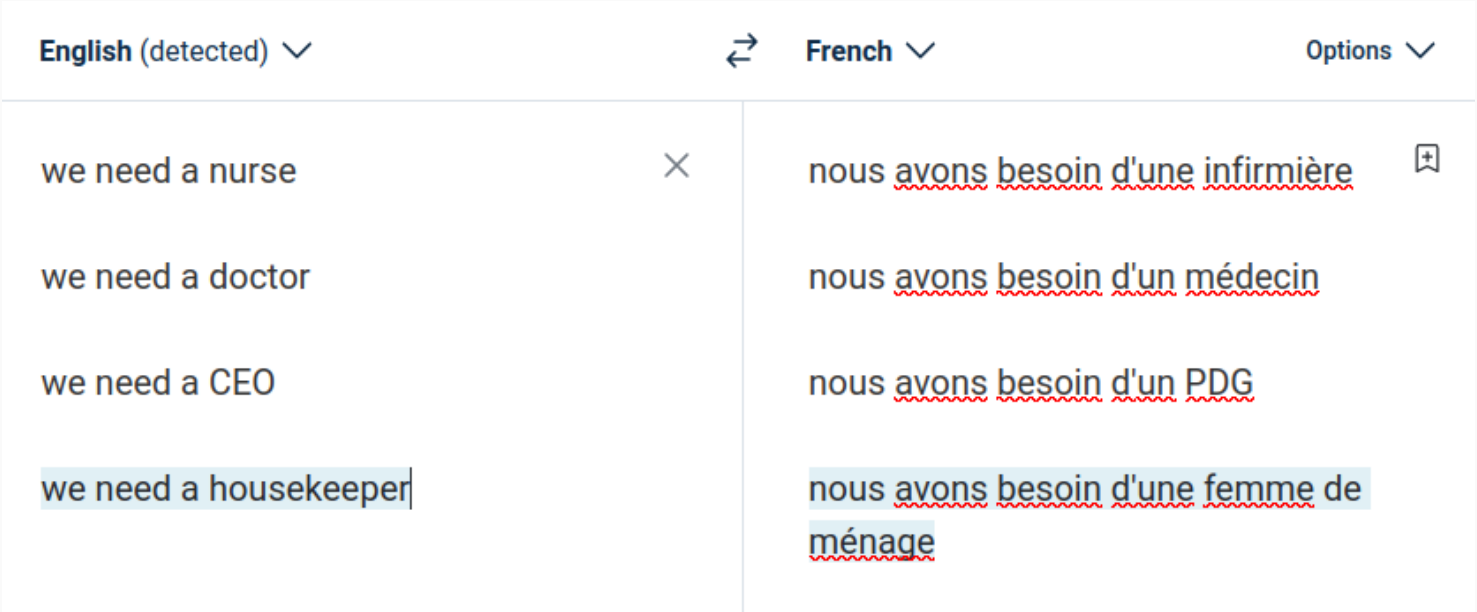
Coronavirus » Plus de 60% des seniors estiment être traités de manière différente, voire injuste à cause de leur âge durant la pandémie de coronavirus. C'est le constat d'une étude menée par la Haute Ecole de travail social de Fribourg. Considérés comme des personnes à risques, les aînés sont particulièrement visés par les mesures prises pour lutter contre le coronavirus. Pourtant, estime le professeur Christian Maggiori, leur voix a été peu entendue. C'est pourquoi l'institution a lancé à la mi-avril un questionnaire en ligne. Les premiers résultats ont été communiqués hier. En l'espace d'une semaine, 2480 personnes de 65 ans et plus venant de toute la Suisse romande (dont 58% de femmes) ont répondu au sondage. L'âge moyen des participants est de 71,8 ans.

Impact des biais de sélection sur les algorithmes d'IA

- Les algorithmes d'IA dépendent de leur base d'entraînement ;
- Tendance à reproduire et amplifier les stéréotypes et à effacer les minorités ;

Impact des biais de sélection sur les algorithmes d'IA

- Les algorithmes d'IA dépendent de leur base d'entraînement ;
- Tendance à reproduire et amplifier les stéréotypes et à effacer les minorités ;
- Exemple sur la traduction automatique :

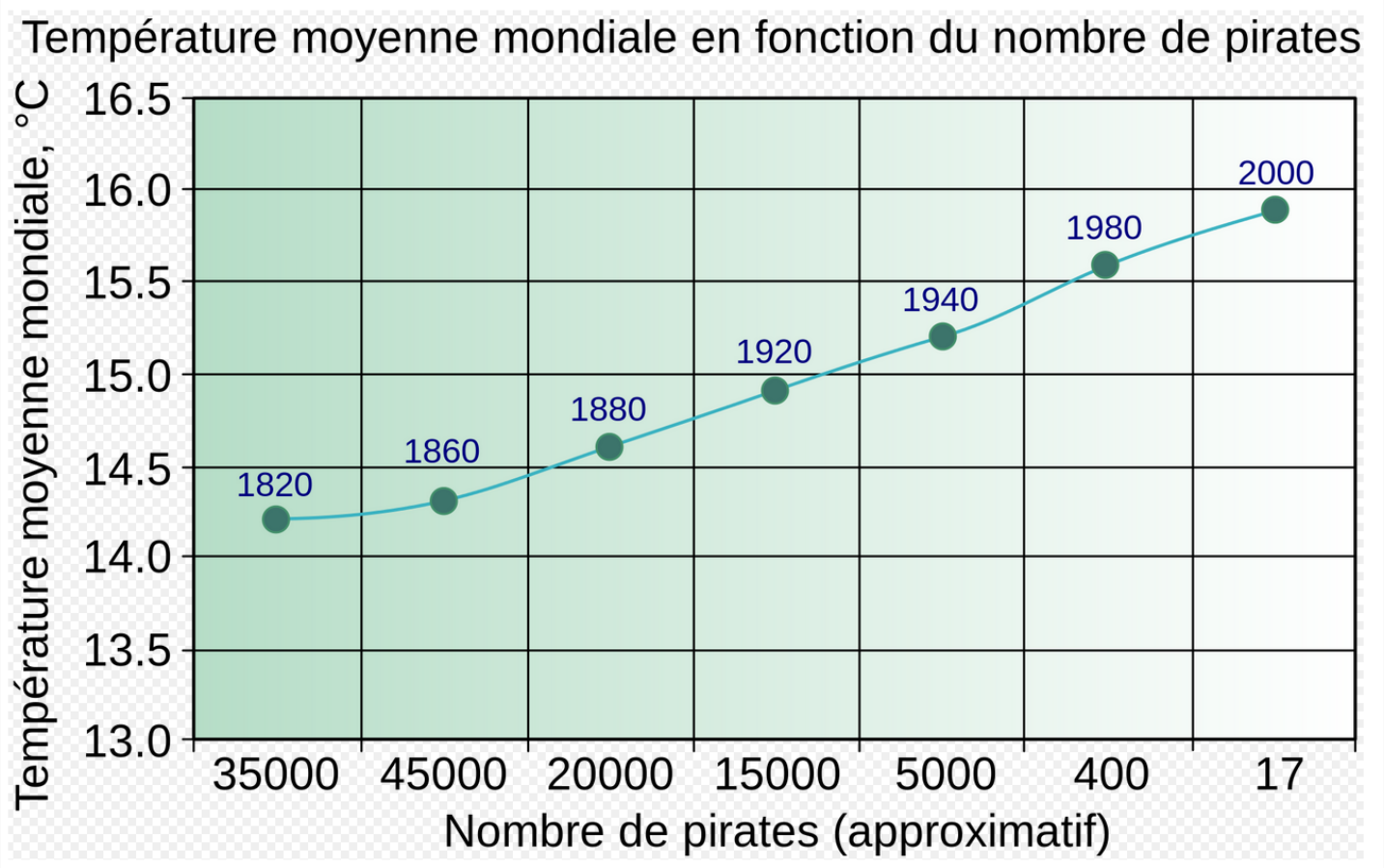


⇒ Il faut être très prudent dans l'utilisation des nouveaux outils (chat GPT, MidJourney,...)

Corrélation, causalité, régression linéaire

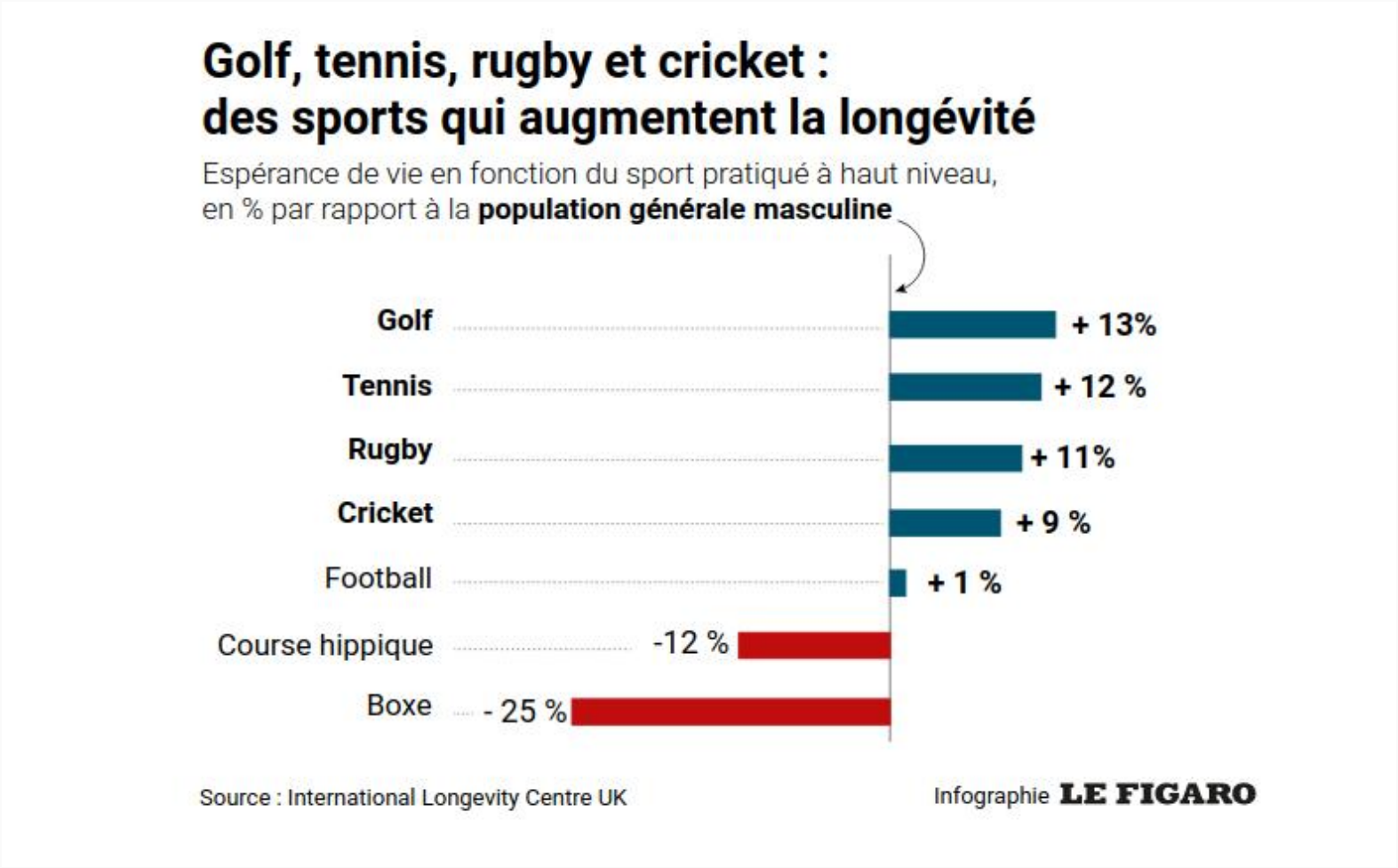
Devenez Pastafaristes

- Les pirates sont d'essence divine ! !



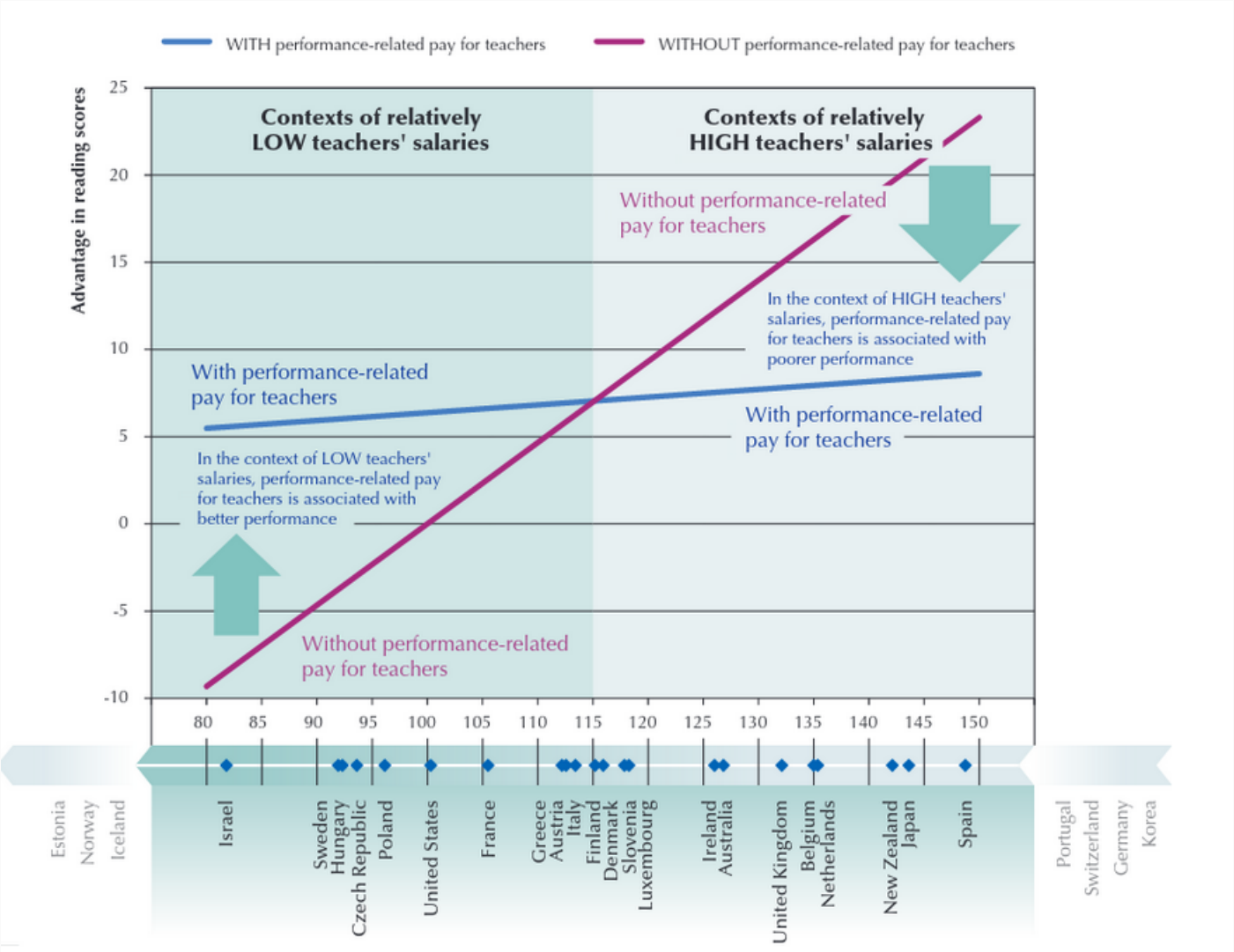
Faites du sport

- Le golf c'est la vie !



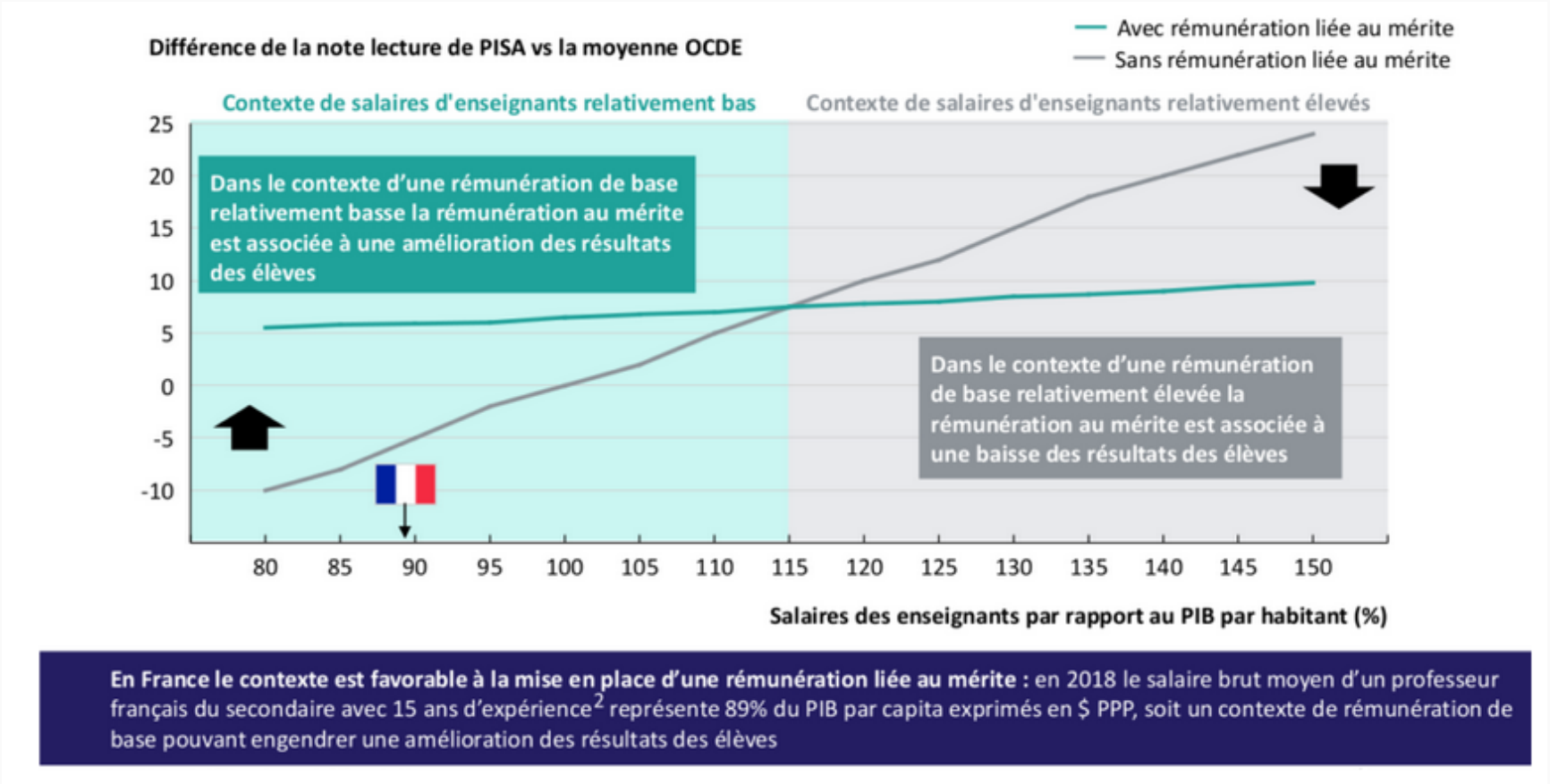
Enseignement et primes au mérite

Le rapport de l'OCDE : résultats des élèves en fonction de la rémunération des enseignants.



Enseignement et primes au mérite

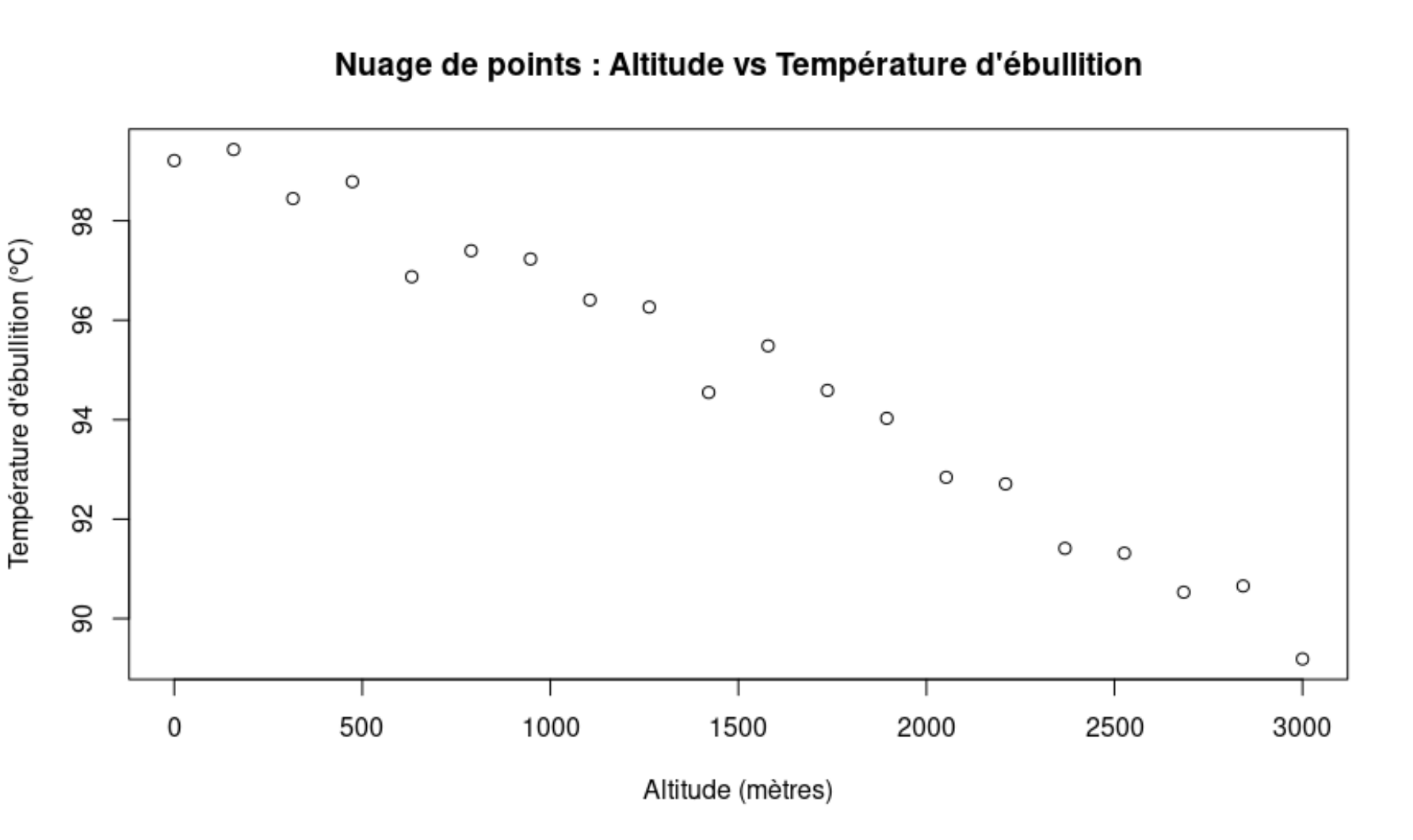
La version McKinsey :



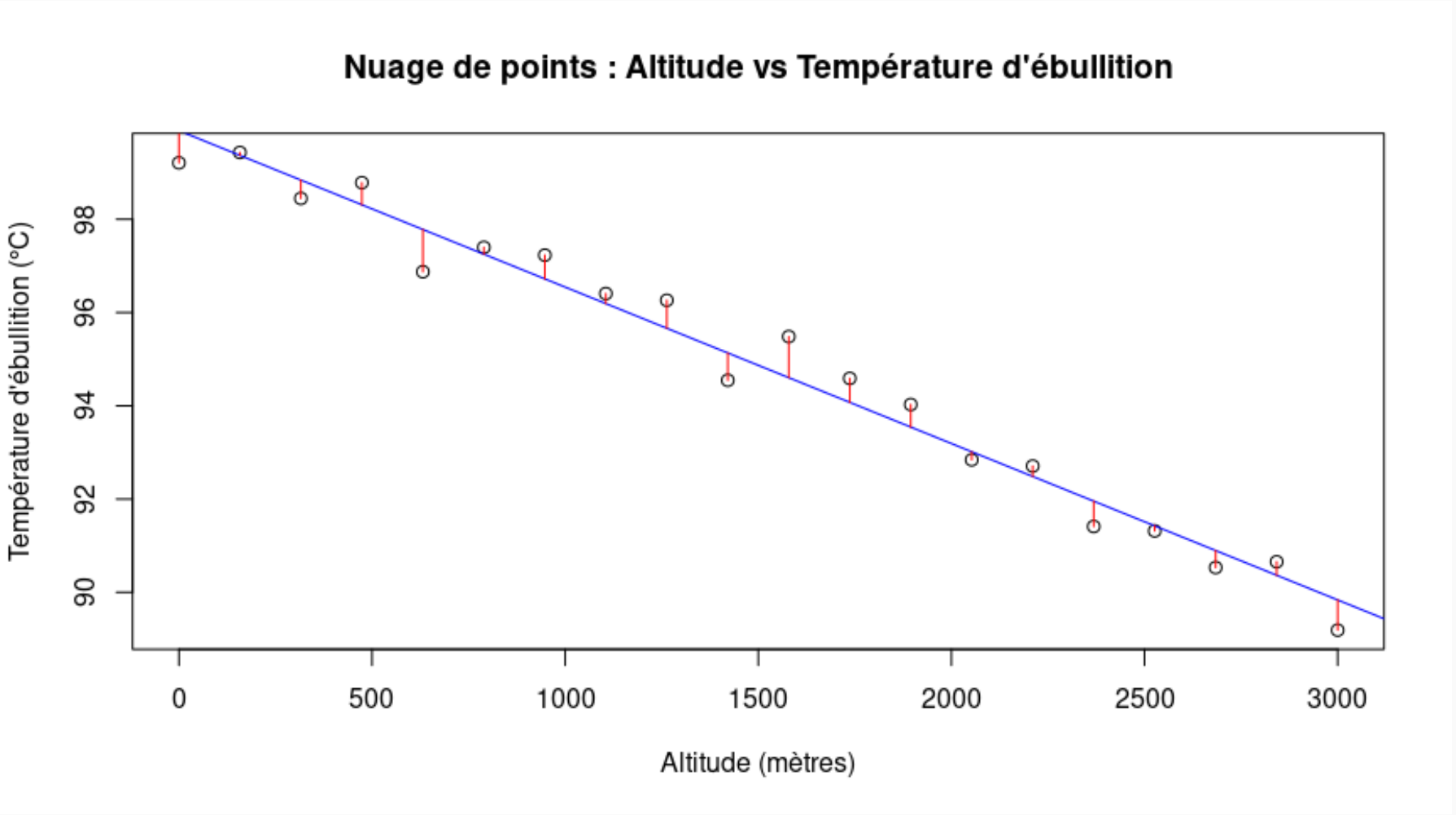
Corrélation et causalité

- Les pièges sont parfois évidents : pirates/température, consommation de chocolat / prix Nobels, ... ;
- Parfois c'est moins intuitif (consommation de tabac/santé) ;
- Il y a des pistes pour estimer/détecter un lien causal.

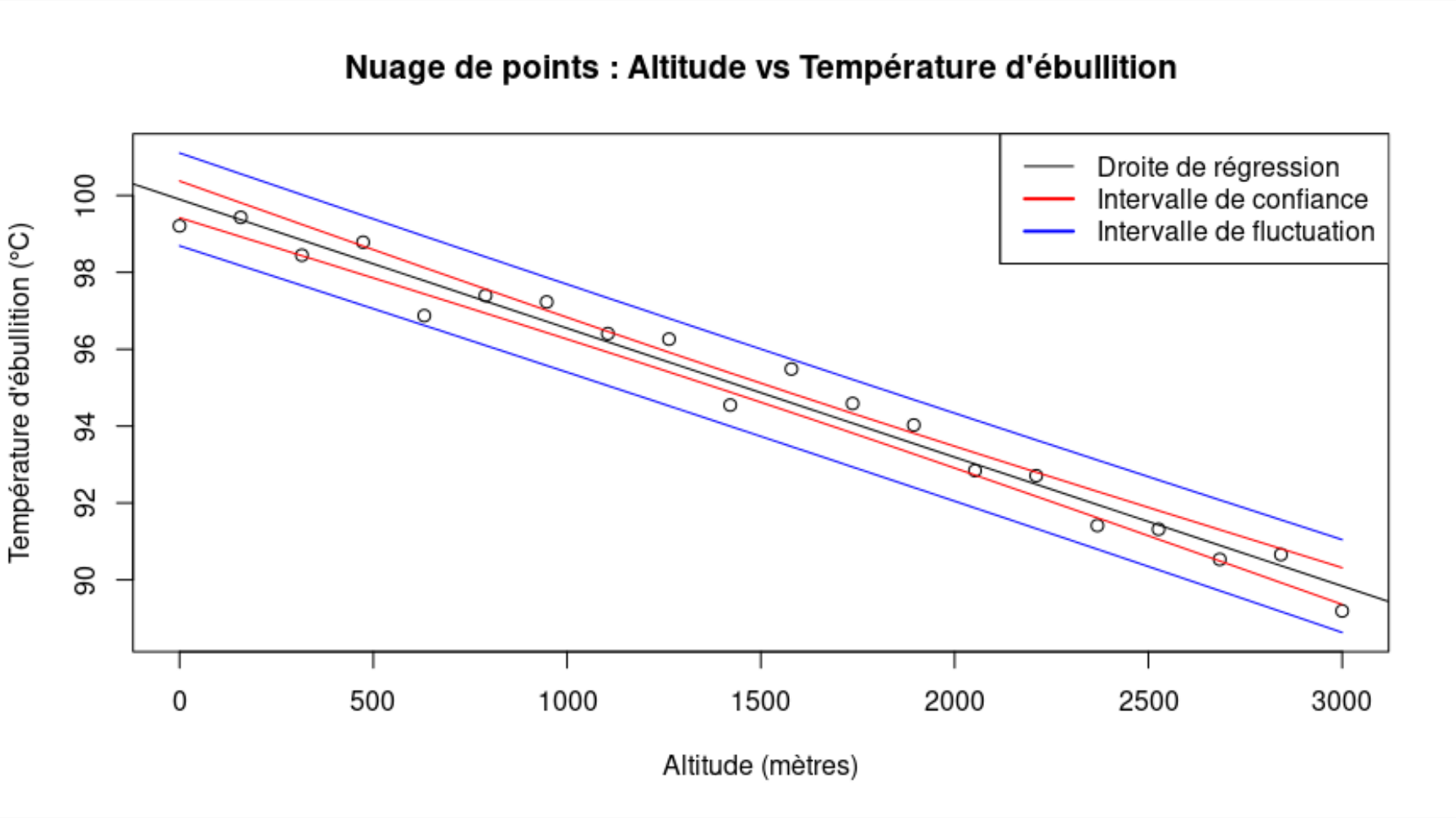
Régression linéaire : problème de départ



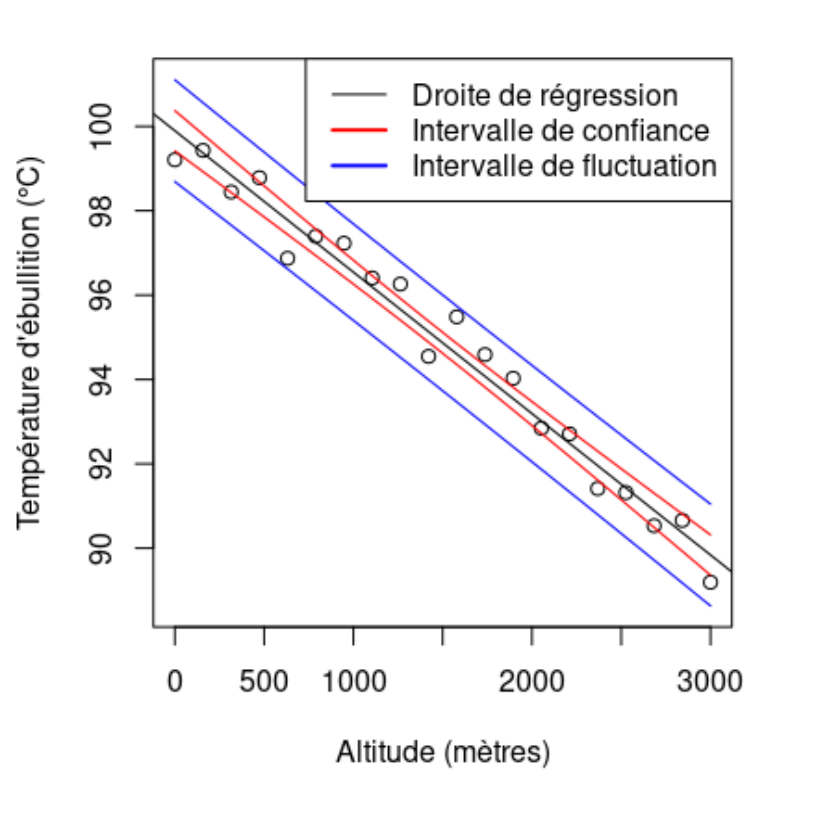
Régression linéaire : méthode des moindres carrés



Régression linéaire : intervalle de confiance et de fluctuation



Régression linéaire : intervalle de confiance et de fluctuation



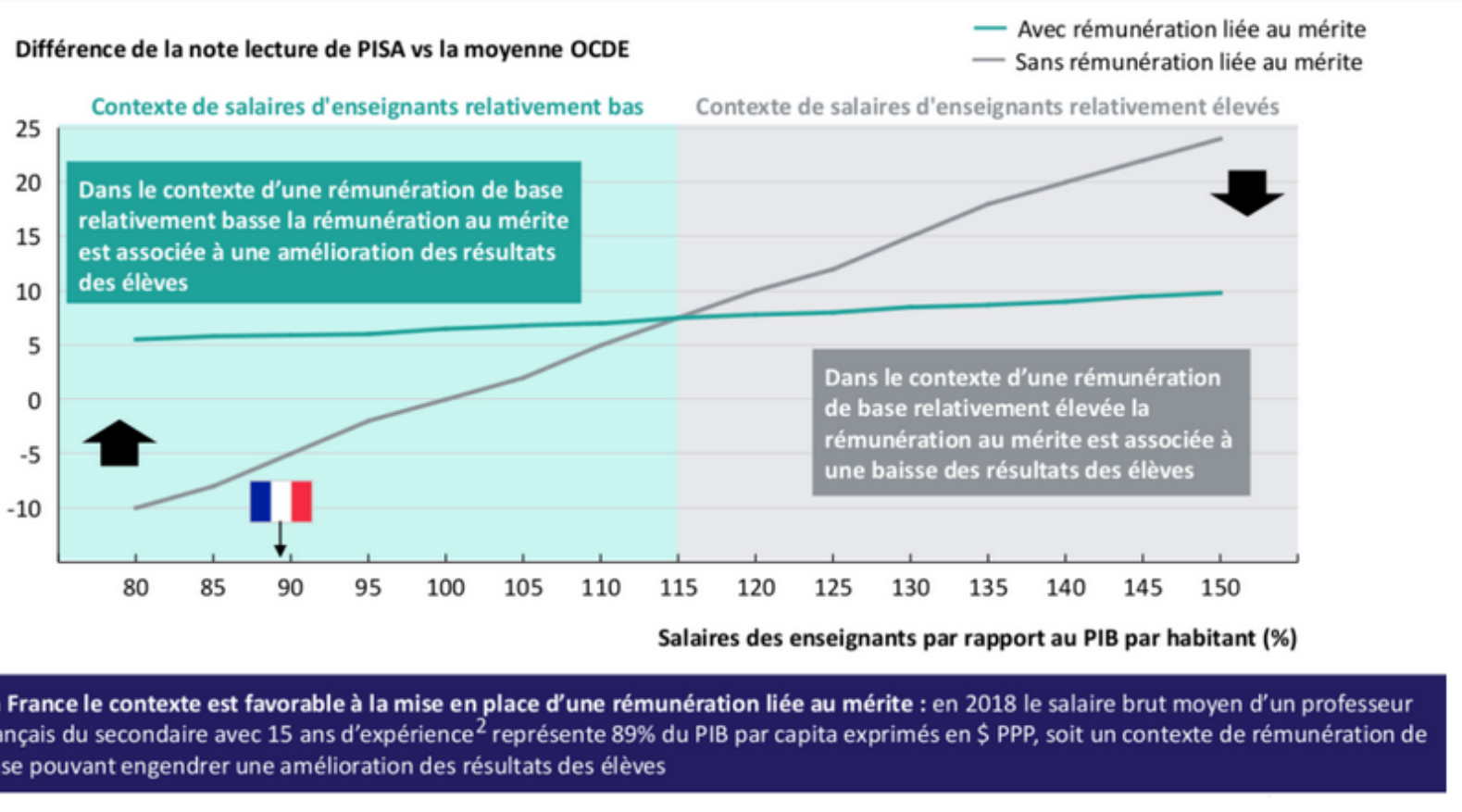
- Intervalle de confiance (95%) : La température moyenne à une altitude donnée a 95% de chances de se trouver dans l'intervalle ;
- Intervalle de fluctuation (95%) : les températures observées ont 95% de chance de se trouver dans l'intervalle.

Régression linéaire : coefficient de corrélation linéaire

- On peut toujours réaliser une régression linéaire ;
- Il faut évaluer sa qualité. Pour cela on dispose d'un outil : le « Coefficient de corrélation linéaire »(ou son carré, noté R^2) ;
- Plus R^2 est proche de 0, moins la relation linéaire est pertinente. Plus il est proche de 1 plus elle l'est ;
- On peut faire des régressions non linéaires ;
- Une corrélation n'est pas une causalité.

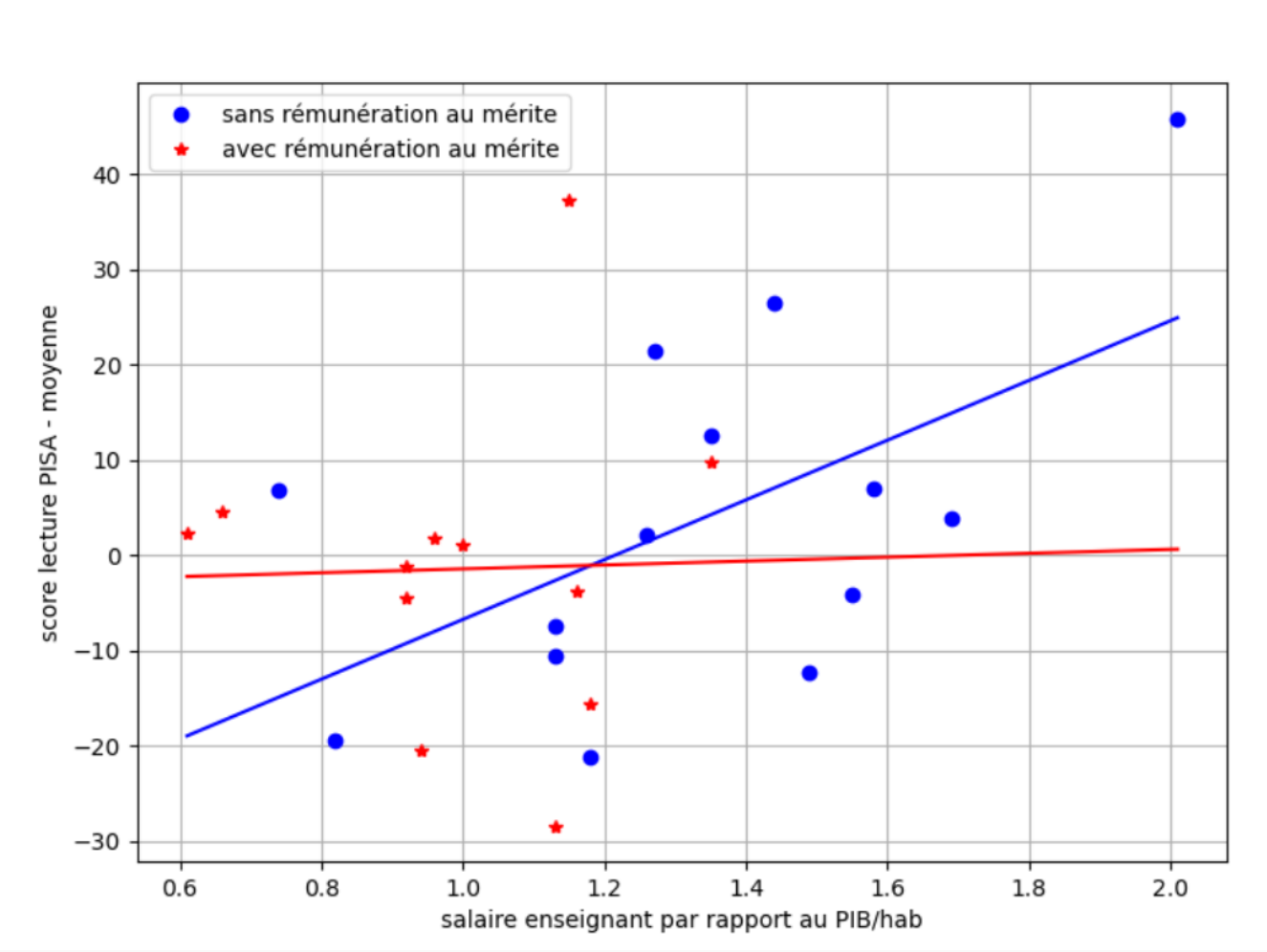
Enseignement et primes au mérite

Rappel de l'exemple :

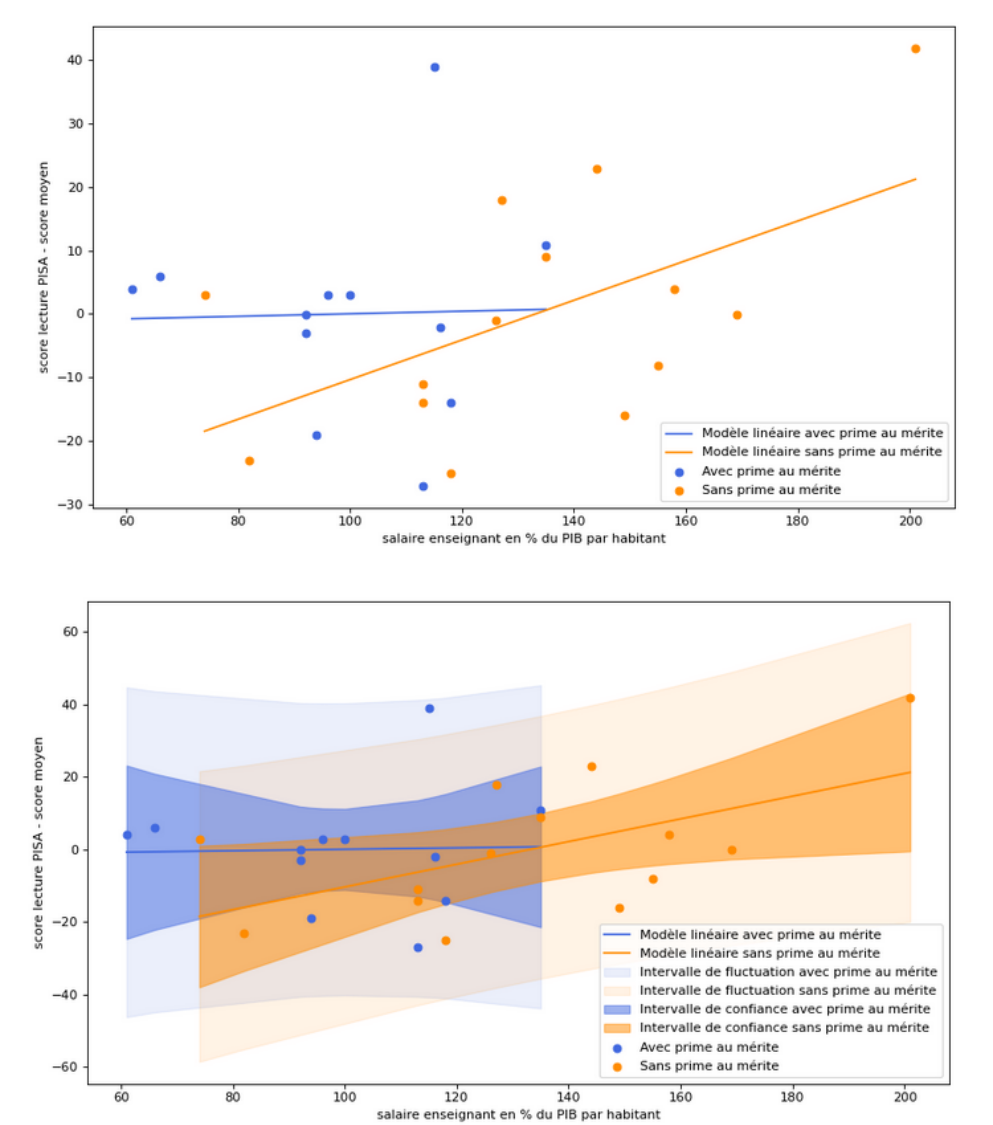


En France le contexte est favorable à la mise en place d'une rémunération liée au mérite : en 2018 le salaire brut moyen d'un professeur français du secondaire avec 15 ans d'expérience[2] représente 89% du PIB par capita exprimés en $ PPP, soit un contexte de rémunération de base pouvant engendrer une amélioration des résultats des élèves

Les données derrière le rapport



Une représentation plus raisonnable



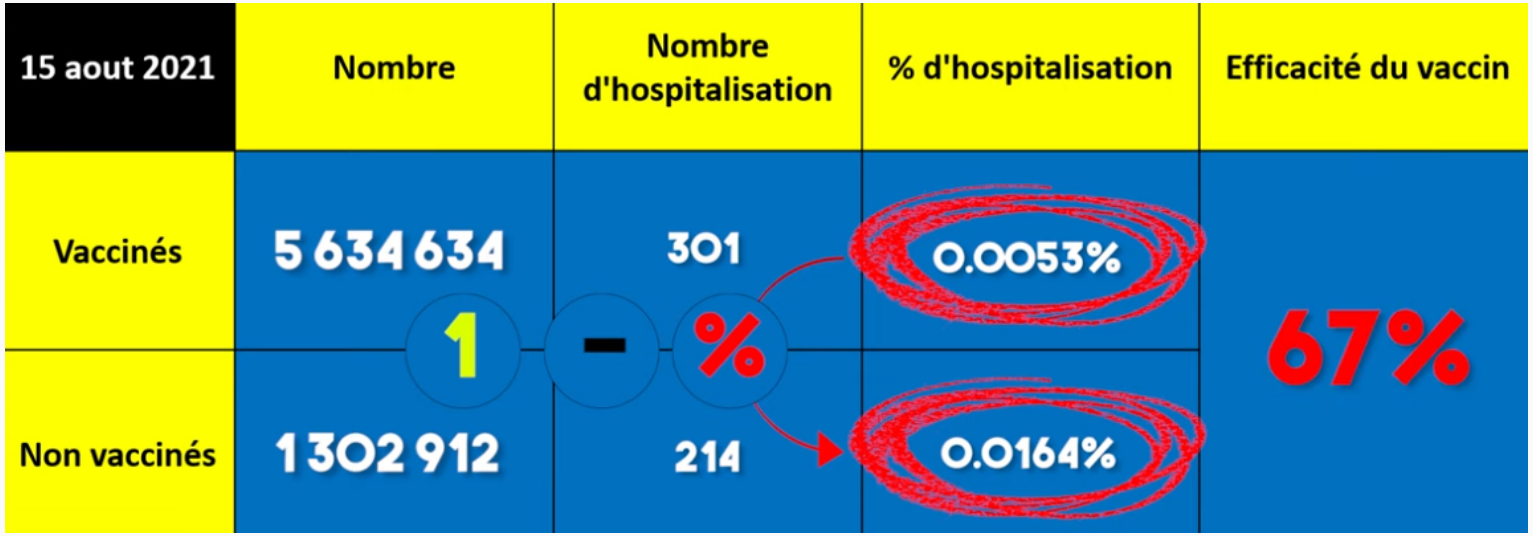
Paradoxe de Simpson (COVID : un vaccin inefficace ?)

Étude israélienne : les chiffres absolus

- Taux de protection annoncé par Pfizer : 0.95 ;
- Été 2021, Israël est un des états ayant le plus vacciné ;
- Étude grandeur nature sur l'efficacité du vaccin.

15 aout 2021	Nombre d'hospitalisation
Vaccinés	301
Non vaccinés	214

Étude israélienne : la première conclusion



Étude israélienne : la première conclusion

15 aout 2021	Nombre	Nombre d'hospitalisation	% d'hospitalisation	Efficacité du vaccin
Vaccinés	5 634 634	301	0.0053%	67%
Non vaccinés	1 302 912	214	0.0164%	

1 - % = 67%

Étude I=israélienne : les résultats par classe d'âge

15 aout 2021	Age	Nombre	Nombre d'hospitalisation	% d'hospitalisation	Efficacité du vaccin
Vaccinés	<50 ANS	3 501 118	11	0.0003%	92%
	>50 ANS	2 133 511	290	0.0136%	85%
Non vaccinés	<50 ANS	1 116 834	43	0.0039%	
	>50 ANS	186 078	171	0.0919%	

Étude israélienne : Au bout des classes d'âge

Age	Non vaccinés	Vaccinés	Non vaccinés hospitalisés	Vaccinés hospitalisés	% hospitalisation non vaccinés	% hospitalisation vaccinés	Efficacité
12-15	383 649	184 549	1	0	0.000261%	0.000000%	100%
16-19	127 745	429 109	2	0	0.001566%	0.000000%	100%
20-29	265 871	991 408	4	0	0.001504%	0.000000%	100%
30-39	194 213	968 837	12	2	0.006179%	0.000206%	97%
40-49	145 355	927 214	24	9	0.016511%	0.000971%	94%
50-59	84 545	747 949	34	22	0.040215%	0.002941%	93%
60-69	65 205	665 717	50	58	0.076681%	0.008712%	89%
70-79	20 512	464 336	39	92	0.190129%	0.019813%	90%
80-89	12 683	208 911	32	100	0.252297%	0.047867%	81%
90+	3 132	46 602	16	18	0.510846%	0.038625%	92%

Étude israélienne : Au bout des classes d'âge

Age	Non vaccinés	Vaccinés	Non vaccinés hospitalisés	Vaccinés hospitalisés	% hospitalisation non vaccinés	% hospitalisation vaccinés	Efficacité
12-15	383 649	184 549	1	0	0.000261%	0.000000%	100%
16-19	127 745	429 109	2	0	0.001566%	0.000000%	100%
20-29	265 871	991 408	4	0	0.001504%	0.000000%	100%
30-39	194 213	968 837	12	2	0.006179%	0.000206%	97%
40-49	145 355	927 214	24	9	0.016511%	0.000971%	94%
50-59	84 545	747 949	34	22	0.040215%	0.002941%	93%
60-69	65 205	665 717	50	58	0.076681%	0.008712%	89%
70-79	20 512	464 336	39	92	0.190129%	0.019813%	90%
80-89	12 683	208 911	32	100	0.252297%	0.047867%	81%
90+	3 132	46 602	16	18	0.510846%	0.038625%	92%

Attention à ne pas aller trop loin !

Étude israélienne : les résultats par classe d'âge

Forçons le trait (données fictives) :

Données fictives	Age	Nombre d'individus	Probabilité d'hospitalisation	Efficacité du vaccin par tranche d'âge	Individus hospitalisés	Efficacité générale
VACCINÉS	<50 ANS	1000	0.5%	90%	0.5	-45%
	>50 ANS	9000	3%	60%	108	
NON VACCINÉS	<50 ANS	9000	0.5%		45	
	>50 ANS	1000	3%		30	

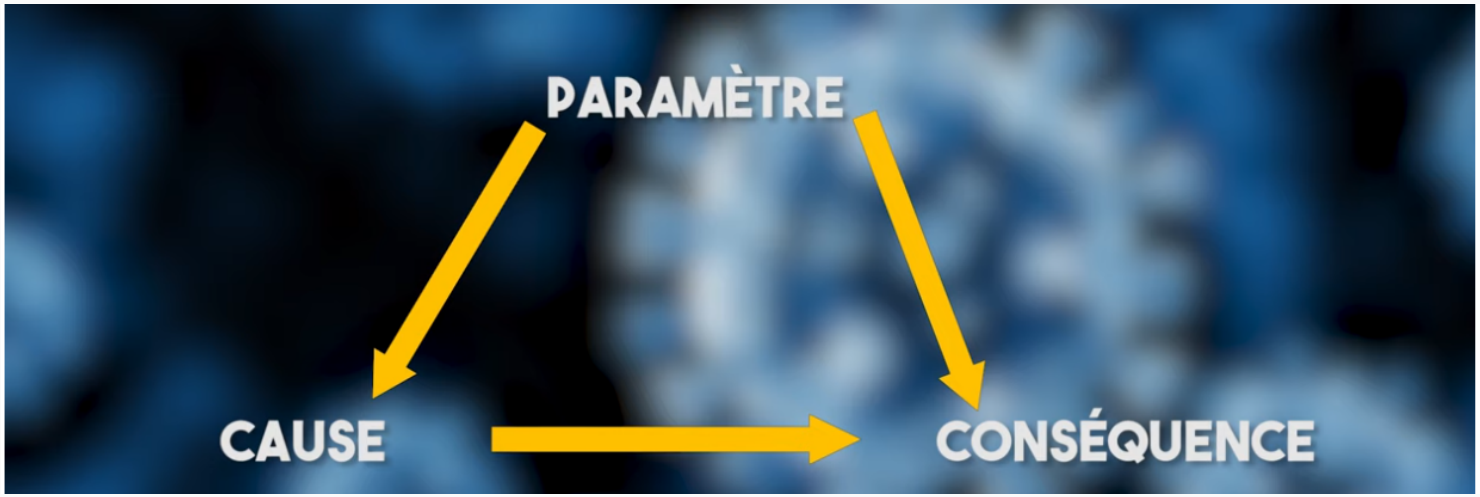
Étude israélienne : les résultats par classe d'âge

Forçons le trait (données fictives) :

Données fictives	Age	Nombre d'individus	Probabilité d'hospitalisation	Efficacité du vaccin par tranche d'âge	Individus hospitalisés	Efficacité générale
VACCINÉS	<50 ANS	1000	0.5%	90%	0.5	-45%
	>50 ANS	9000	3%	60%	108	
NON VACCINÉS	<50 ANS	9000	0.5%		45	
	>50 ANS	1000	3%		30	

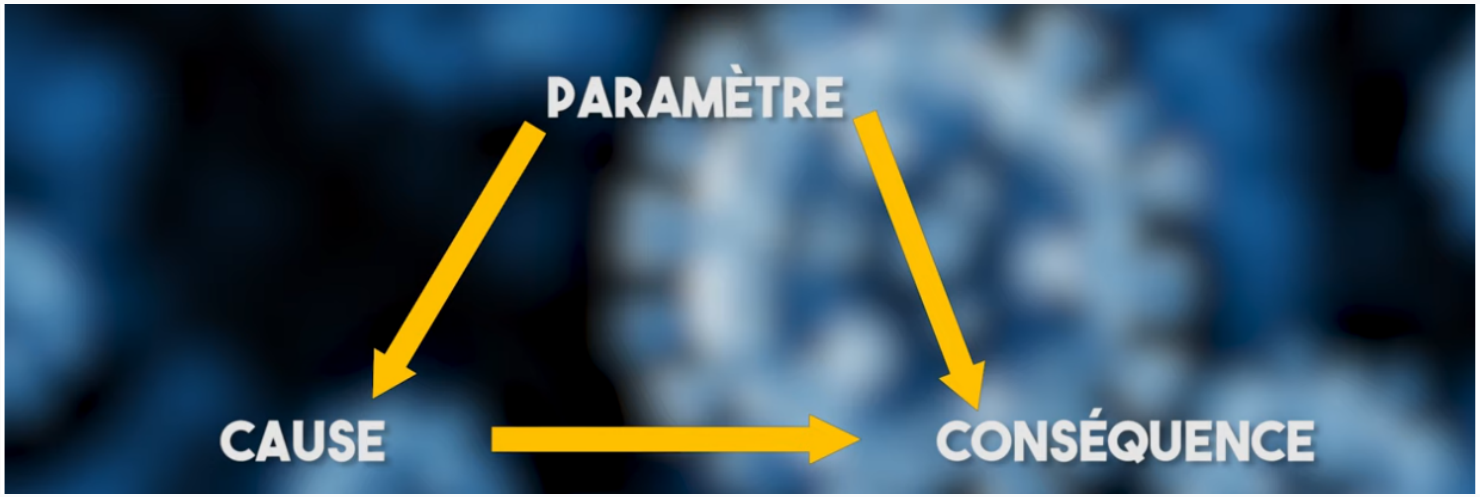
Le paradoxe de Simpson

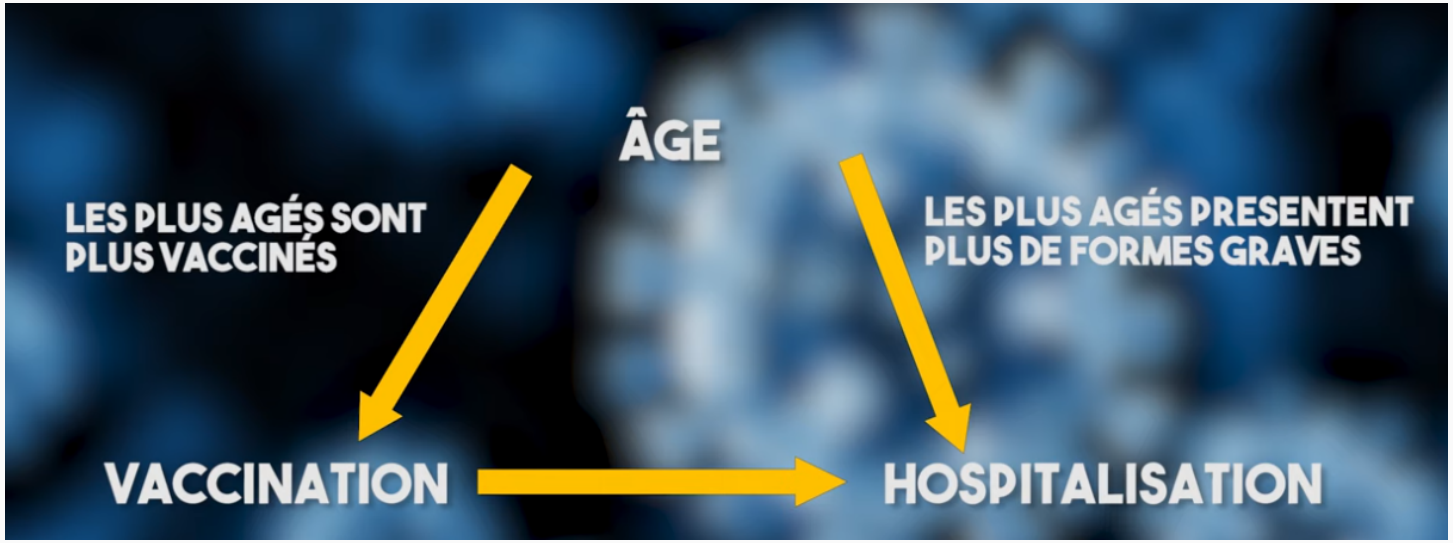
Origine : les facteurs de confusion



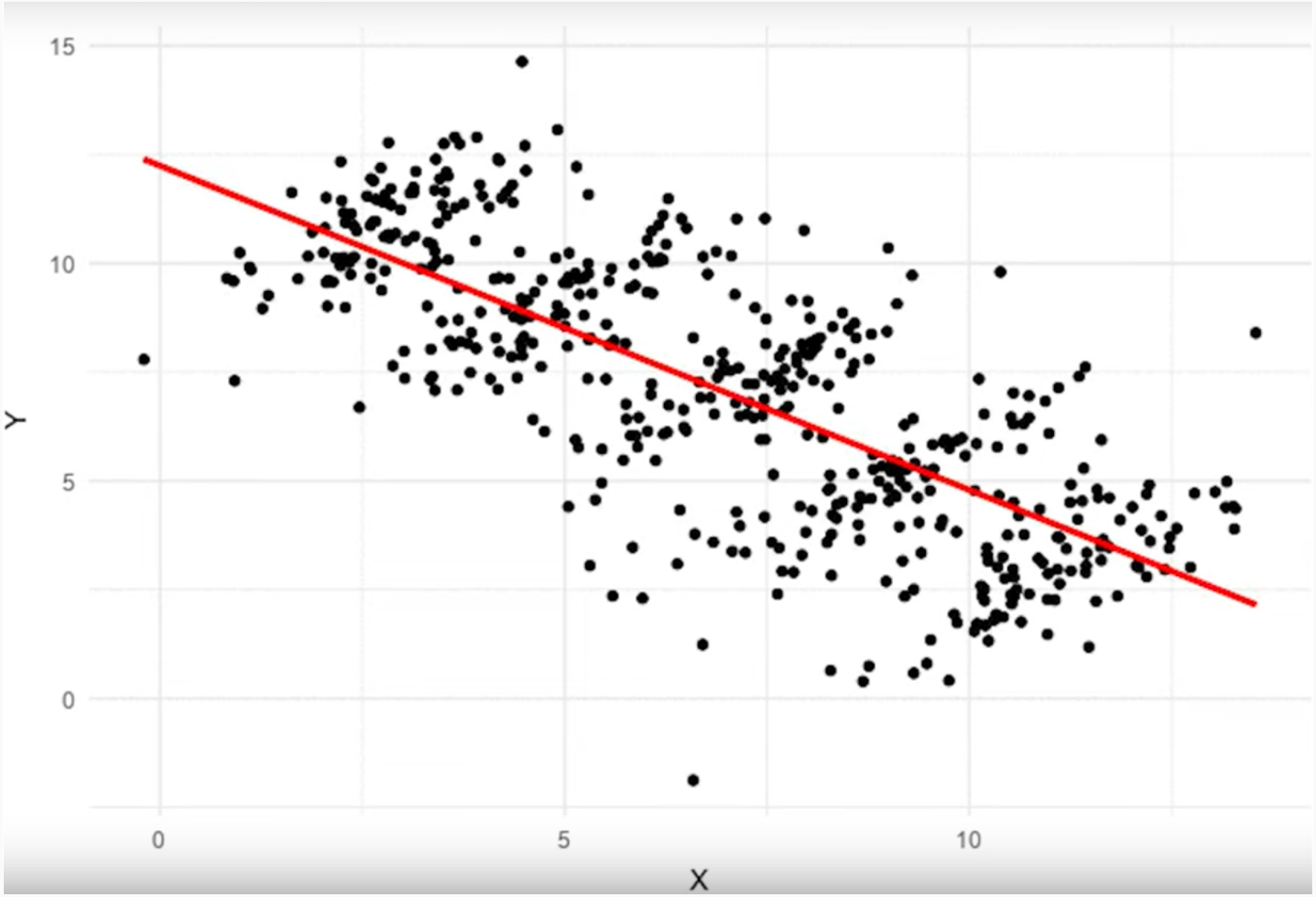
Le paradoxe de Simpson

Origine : les facteurs de confusion

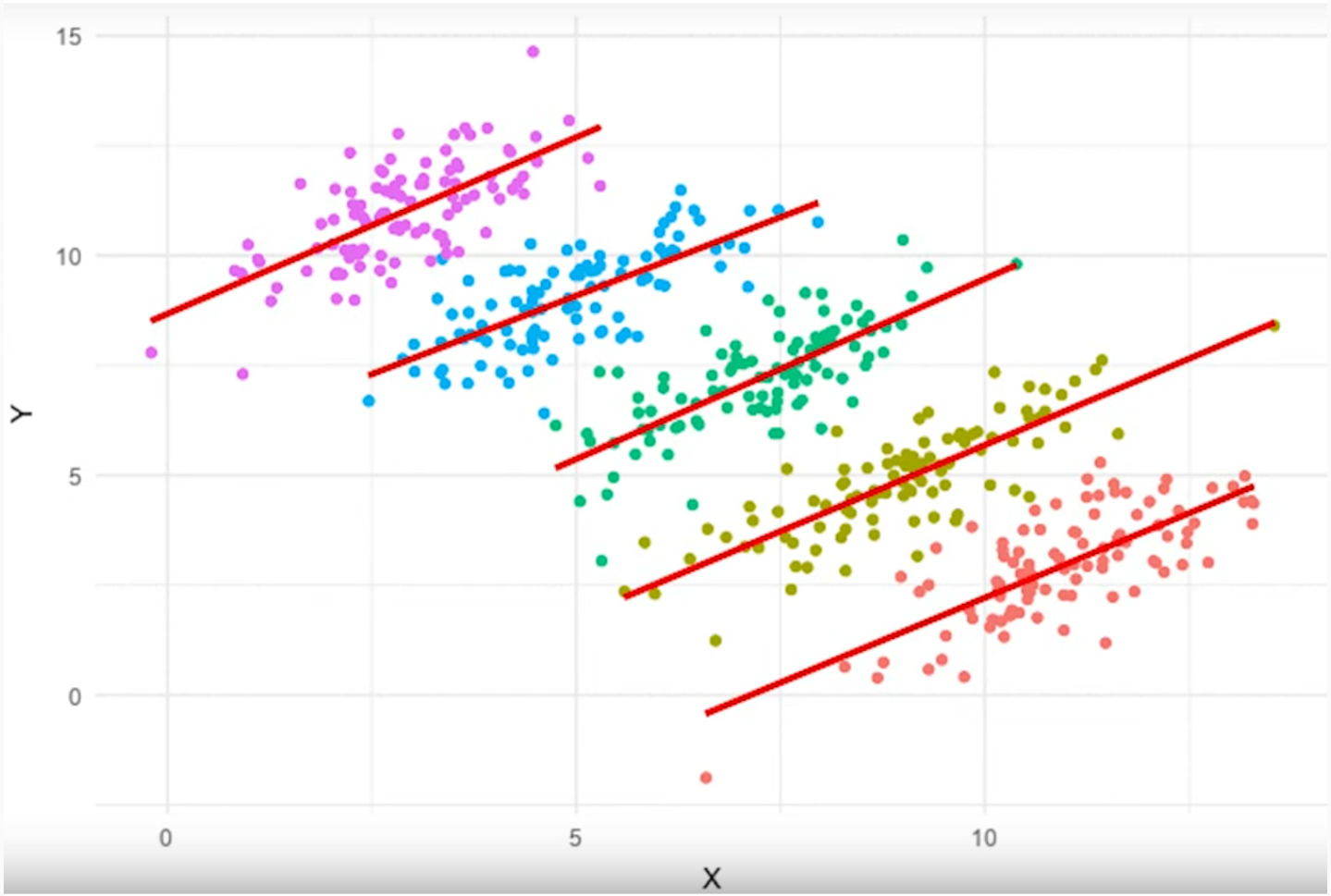




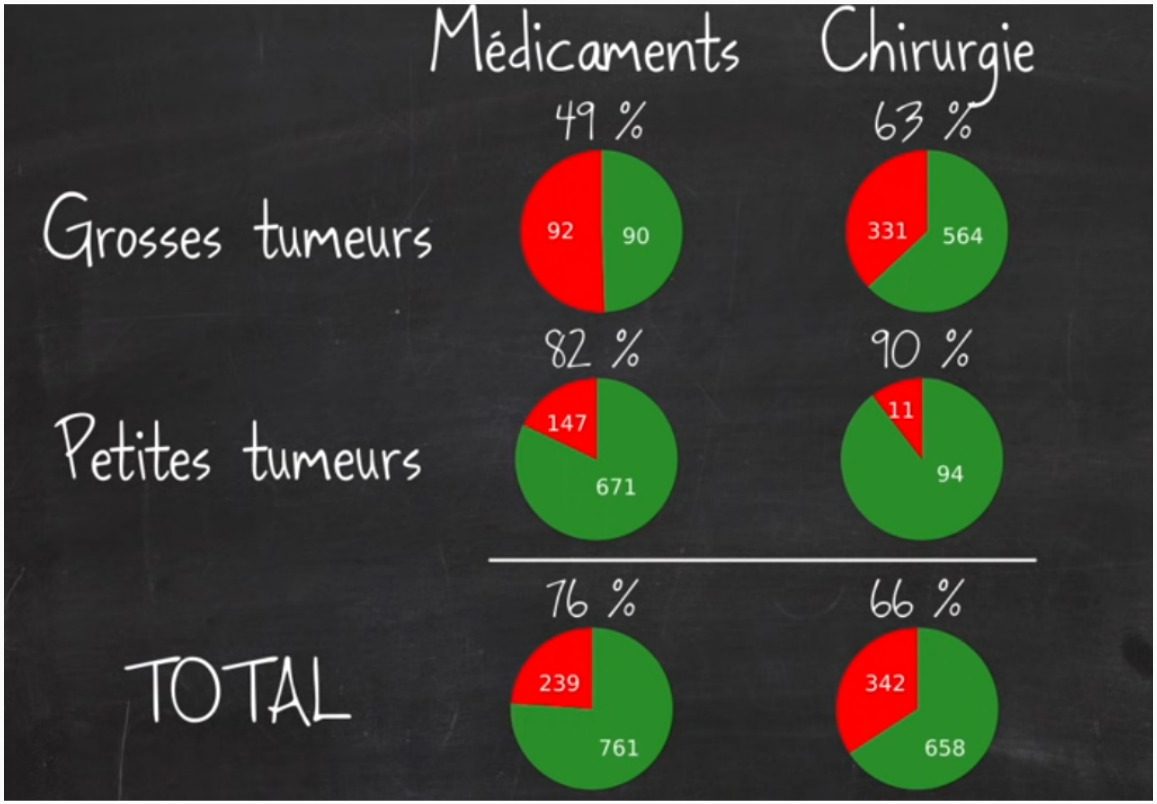
Le paradoxe de Simpson



Le paradoxe de Simpson



Le paradoxe de Simpson



Le paradoxe de Simpson

Comment on l'évite ?

- Etude prospective (vs rétrospective) : randomisation de l'échantillon ;
- Élimination des facteurs de confusion par stratification...

  $\rightarrow$ Problème : ne pas en oublier.

Éviter Le paradoxe de Simpson

Pour l'étude sur les vaccins :

- L'âge est un facteur de confusion évident ;
- Les co-morbidités ?
- La catégorie socio-professionnelle ?
- ...

Pour l'étude sur les tumeurs :

- La taille de la tumeur est un facteur de confusion évident ;
- La consommation de tabac ?
- ... ;

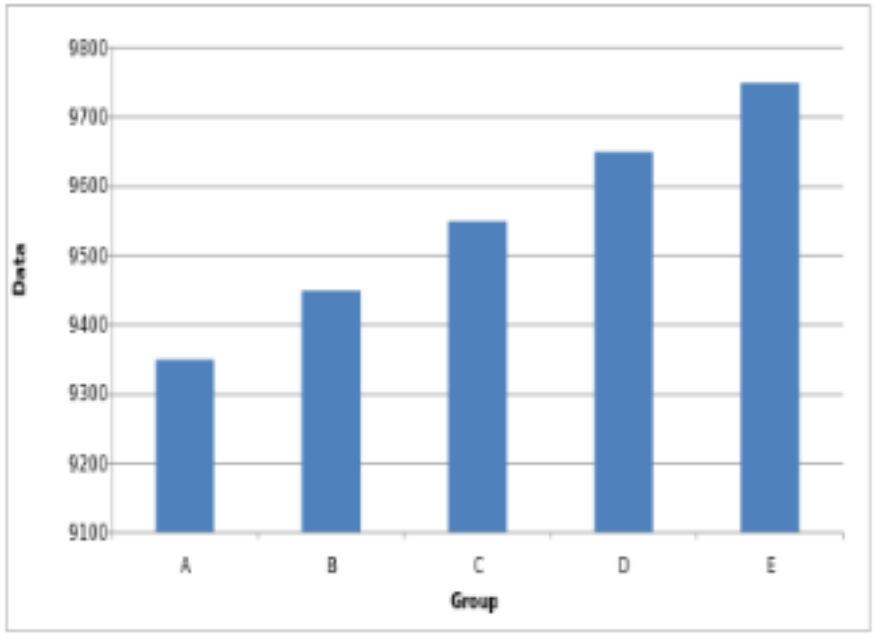
Comment mentir avec des graphiques ?

1. Choisissez votre échelle

- Les diapositives de cette partie sont empruntées soit à C. Bontemps (`http://data.visualisation.free.fr/`) soit au collectif « Cortecs »(`https://cortecs.org/`.

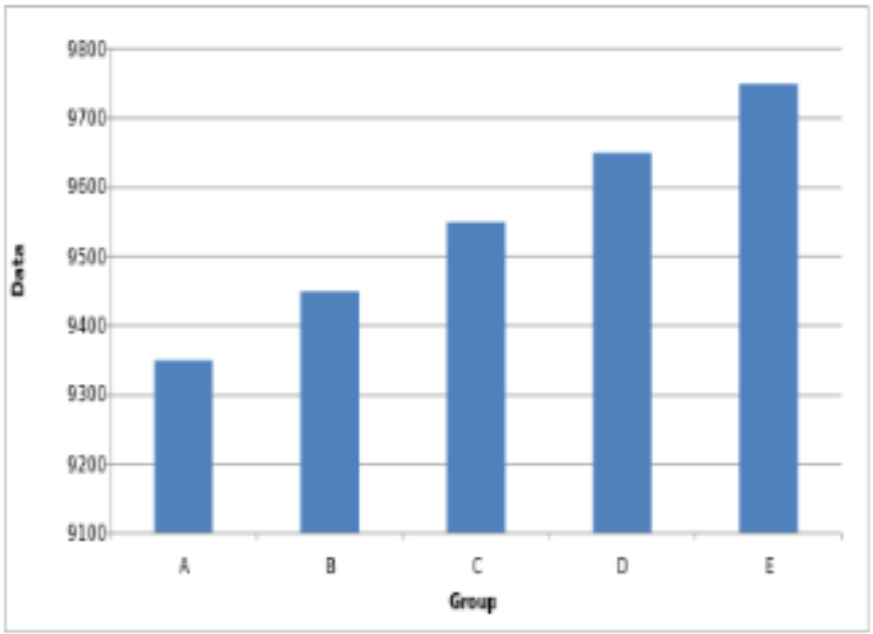
1. Choisissez votre échelle

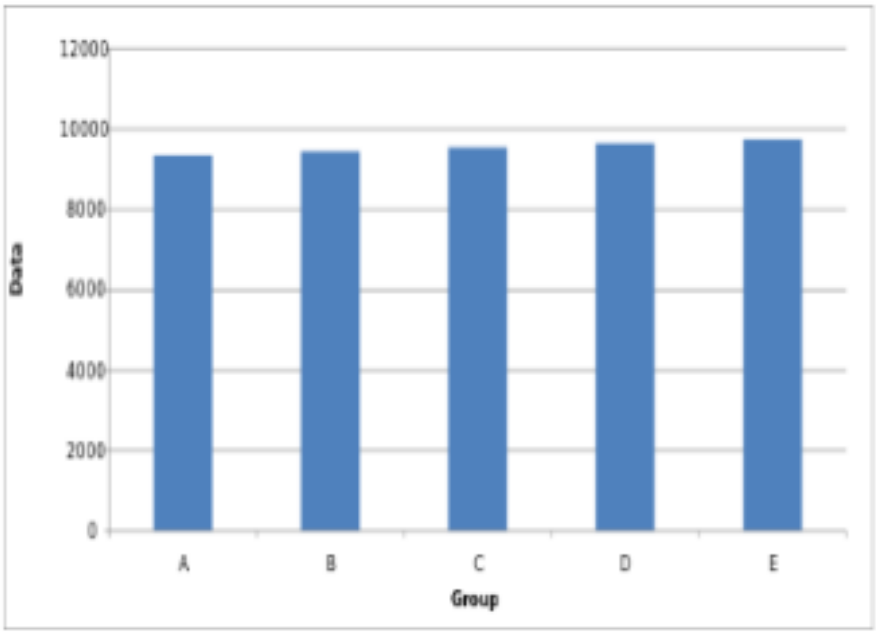
- Les diapositives de cette partie sont empruntées soit à C. Bontemps (`http://data.visualisation.free.fr/`) soit au collectif « Cortecs »(`https://cortecs.org/`.



1. Choisissez votre échelle

- Les diapositives de cette partie sont empruntées soit à C. Bontemps (`http://data.visualisation.free.fr/`) soit au collectif « Cortecs »(`https://cortecs.org/`.

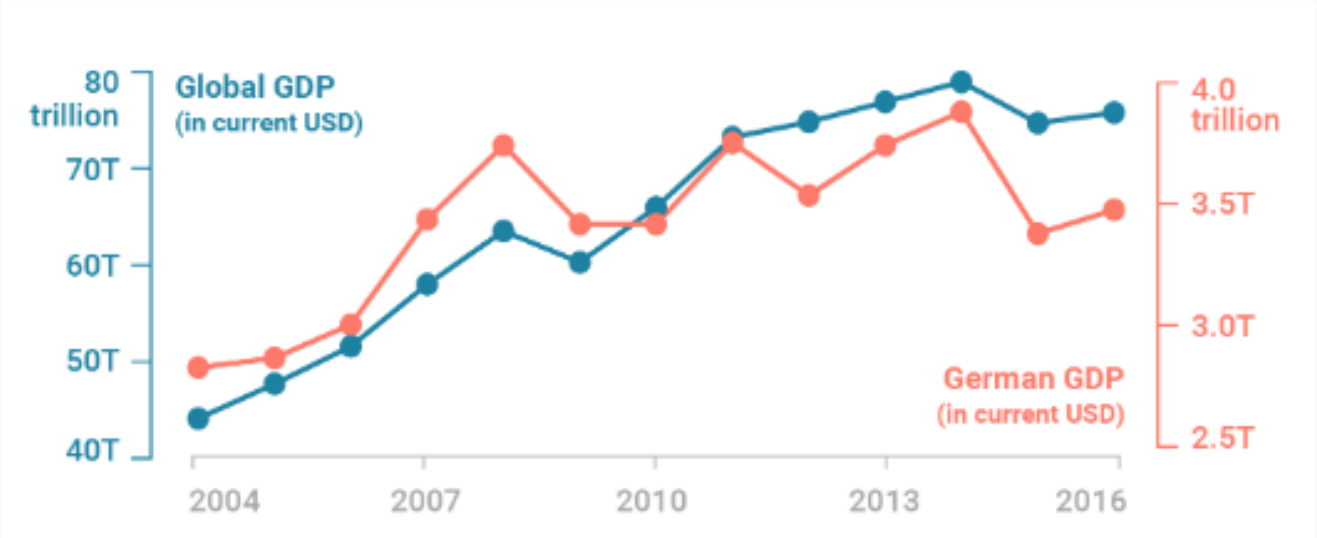




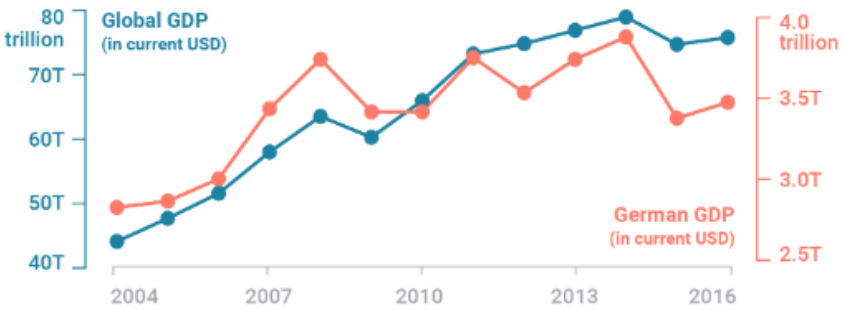
1. Choisissez votre échelle

1. Choisissez votre échelle

2. Prenez deux axes

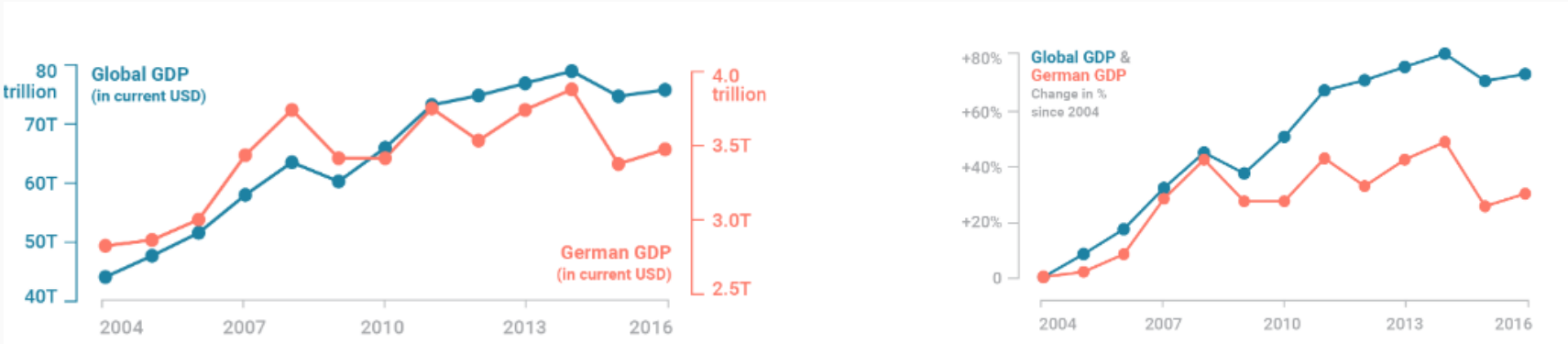


2. Prenez deux axes



2. Prenez deux axes

La solution : passer aux indices !



Exemple : le CAC40 démarre à un indice de base 1000 au 31 décembre 1987.

3. Choisissez votre cadre

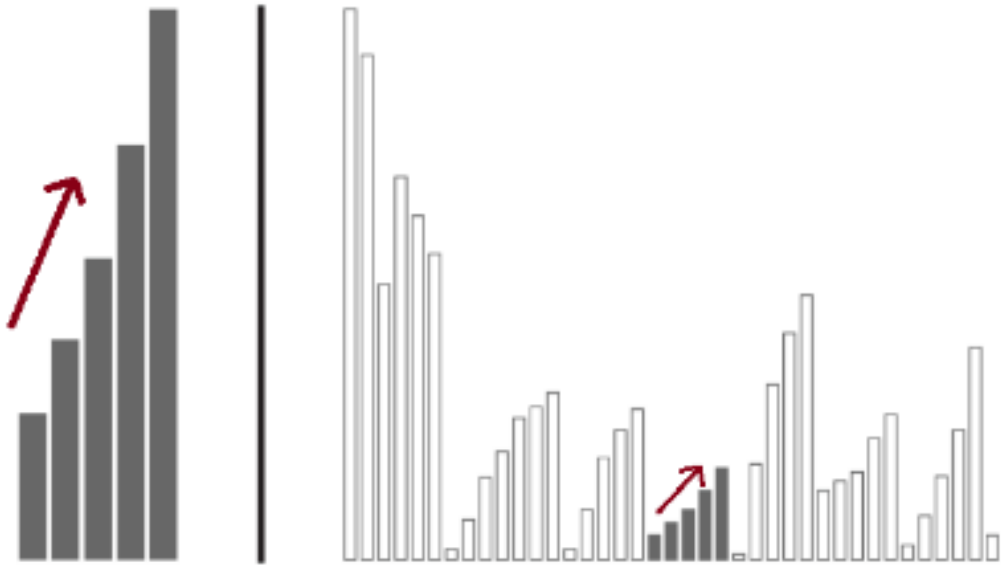


FIGURE – Cela s'appelle le "*Cherry picking*" !

3. Choisissez votre cadre

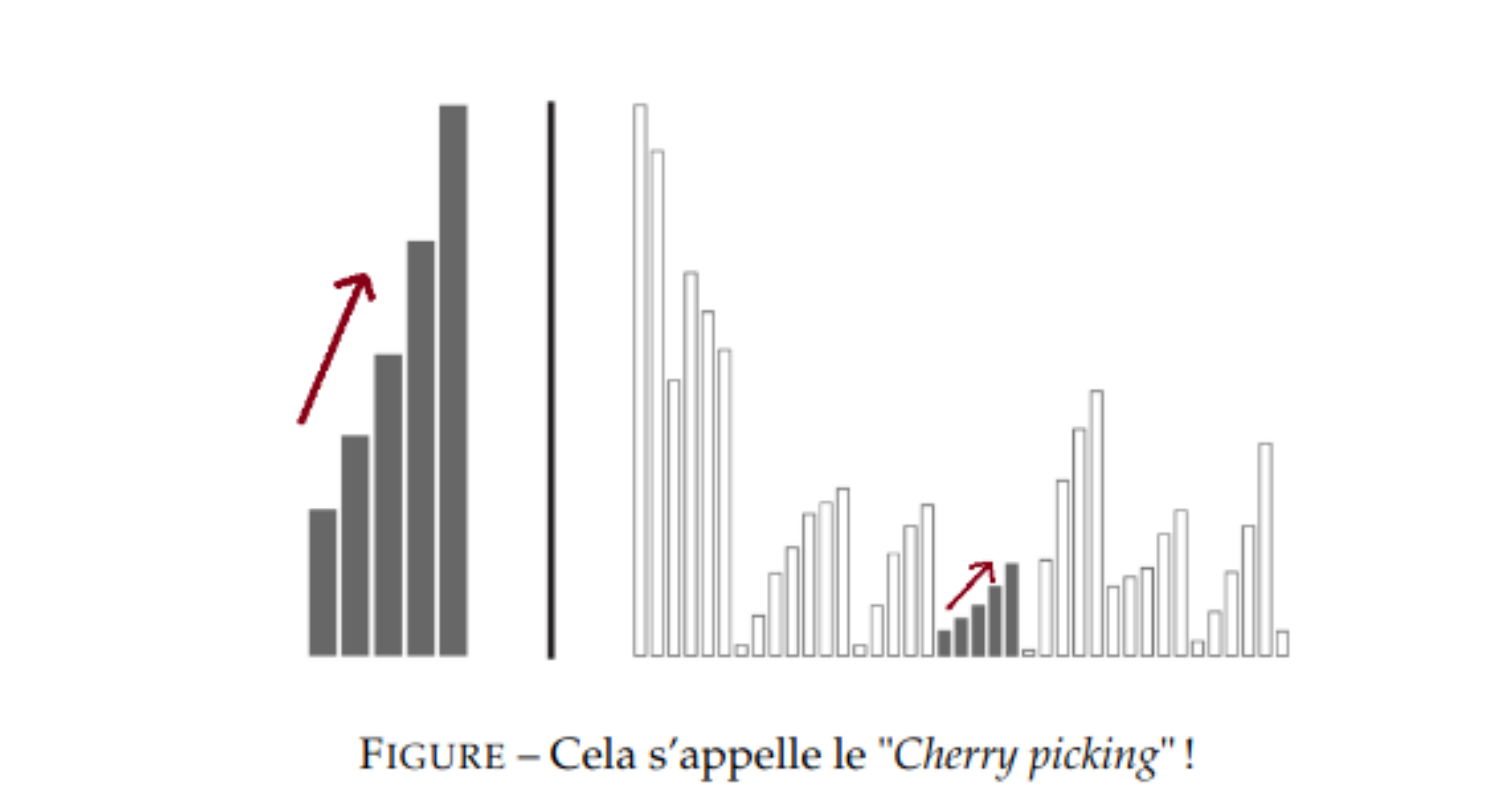


FIGURE – Cela s'appelle le "*Cherry picking*" !

Oui mais personne ne fait ça ...

3. Choisissez votre cadre

... Sauf les JT

Un classique des JT (TF1) :

FIGURE – Que compare-t-on ?

3. Choisissez votre cadre

Quelles sont les données réellement ?

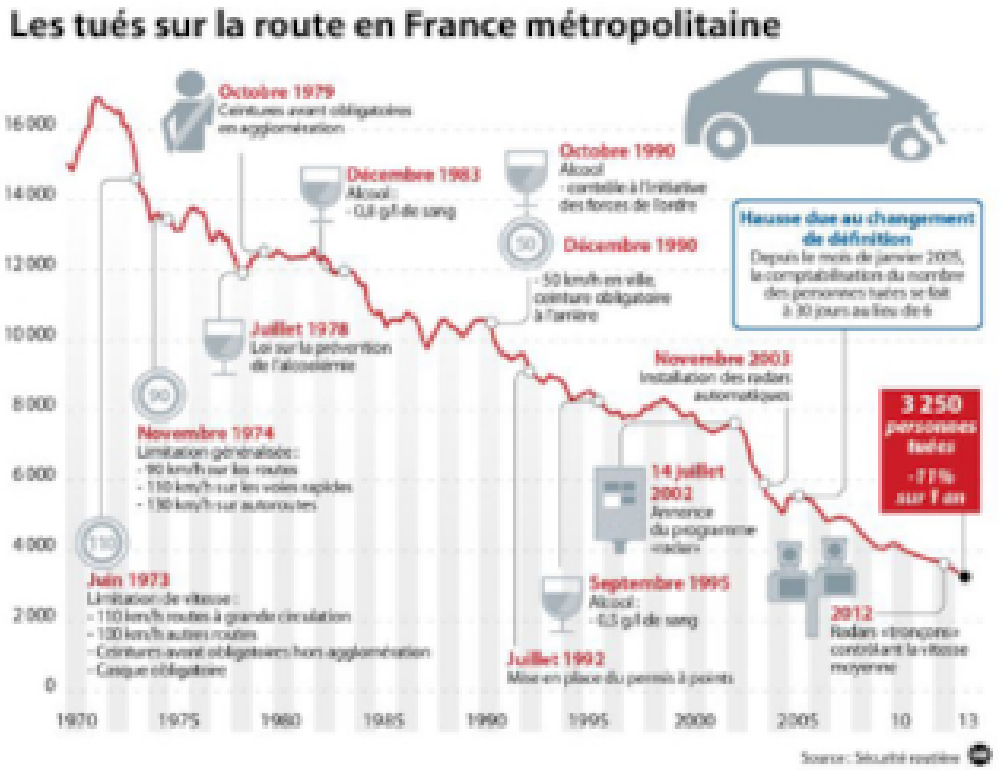
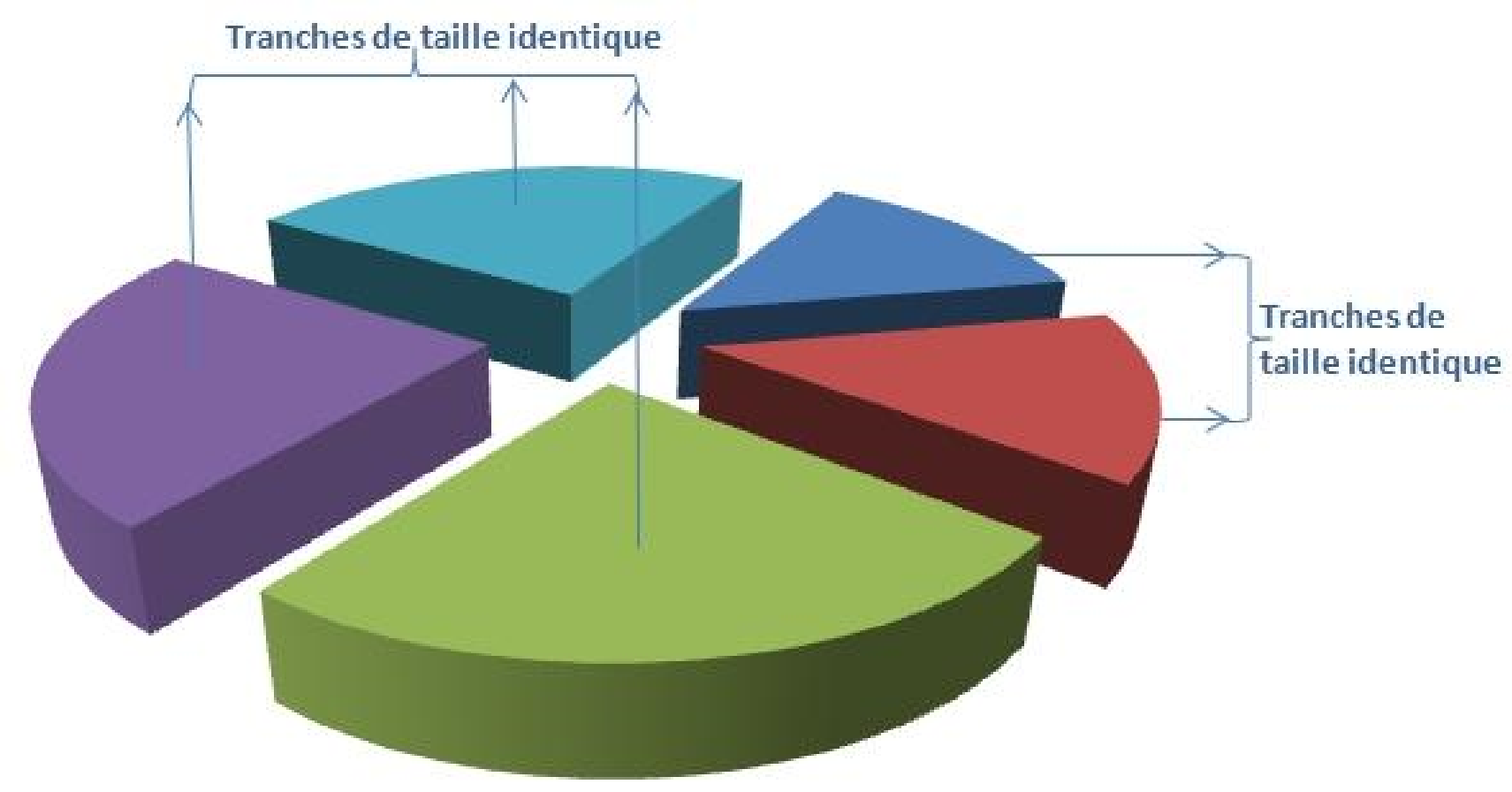


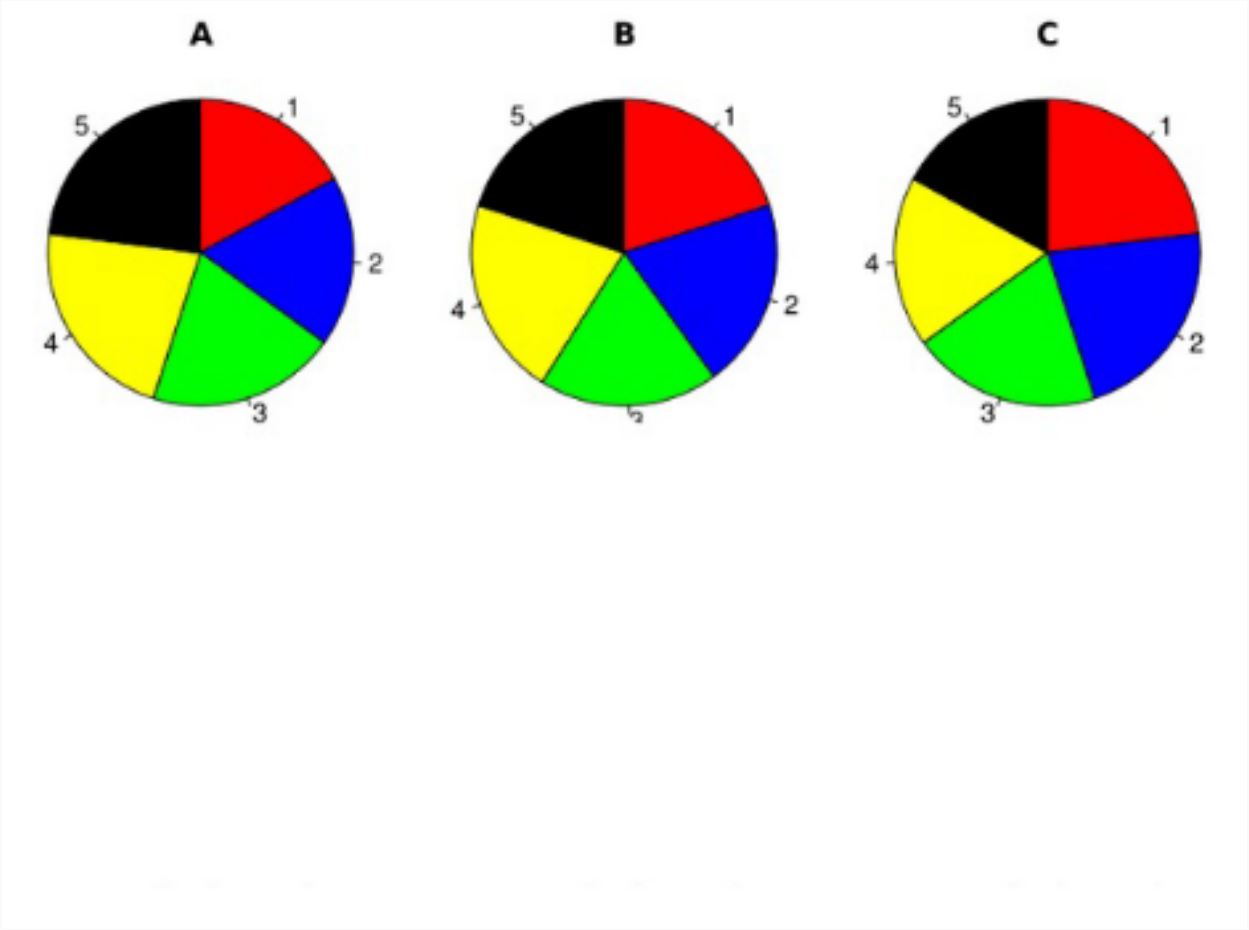
FIGURE – Accidents 1970-2013

4. Vive les camemberts en 3D

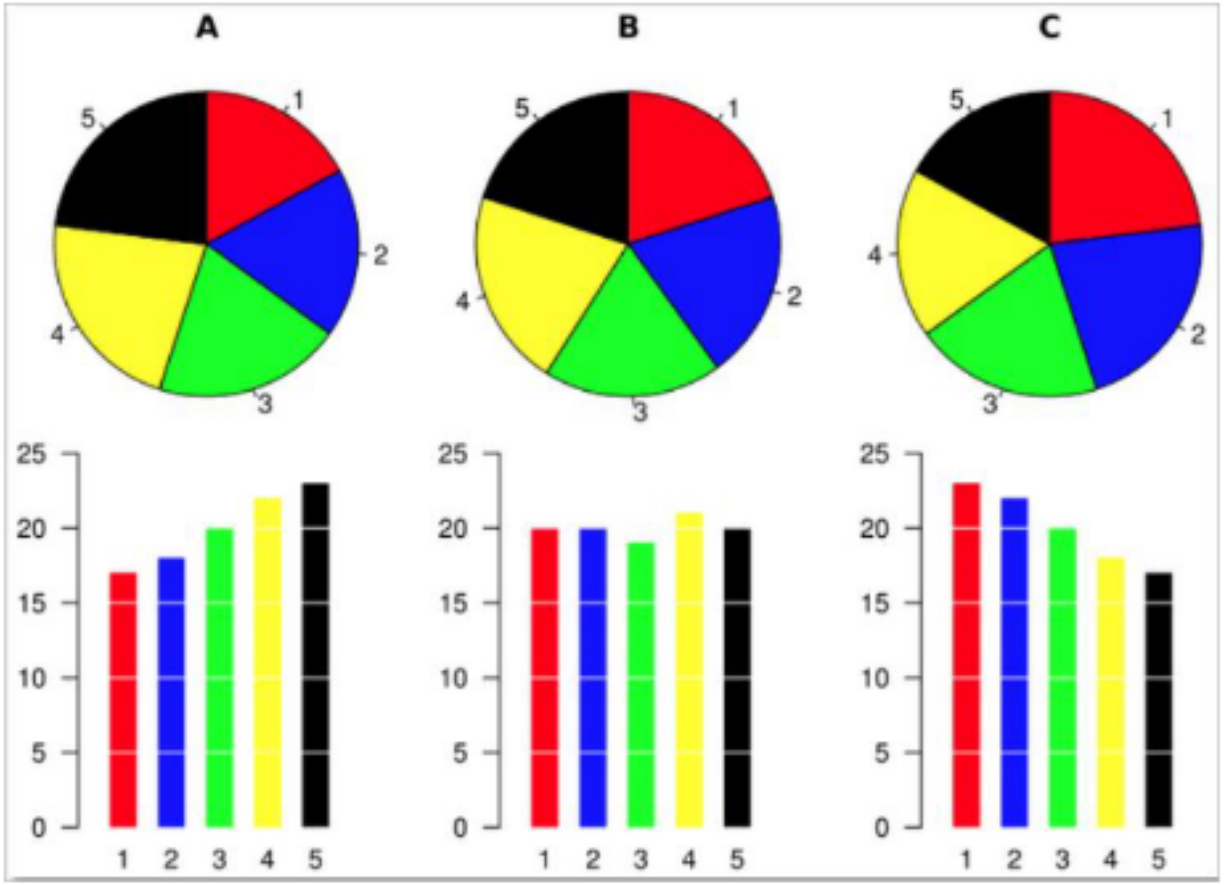


4b. Et en 2D ?

Saurez-vous dire quelle quartier est le plus grand ?



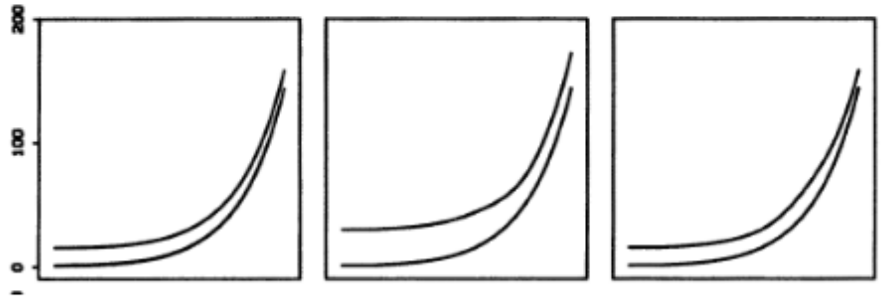
4b. Et en 2D ?



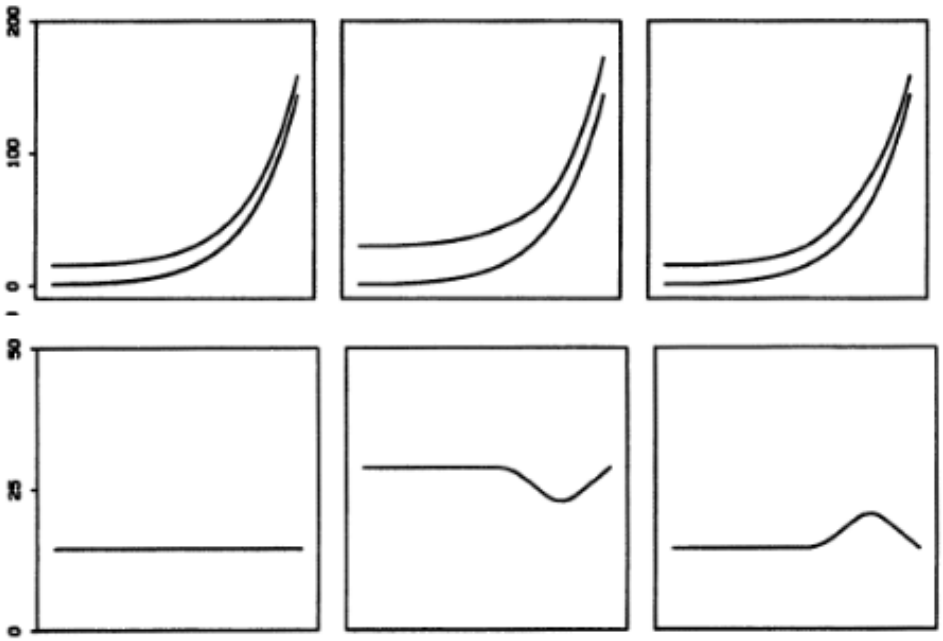
L'oeil humain évalue très mal les surfaces !

5. Comparez des courbes

Pouvez-vous évaluer l'écart entre les courbes ?

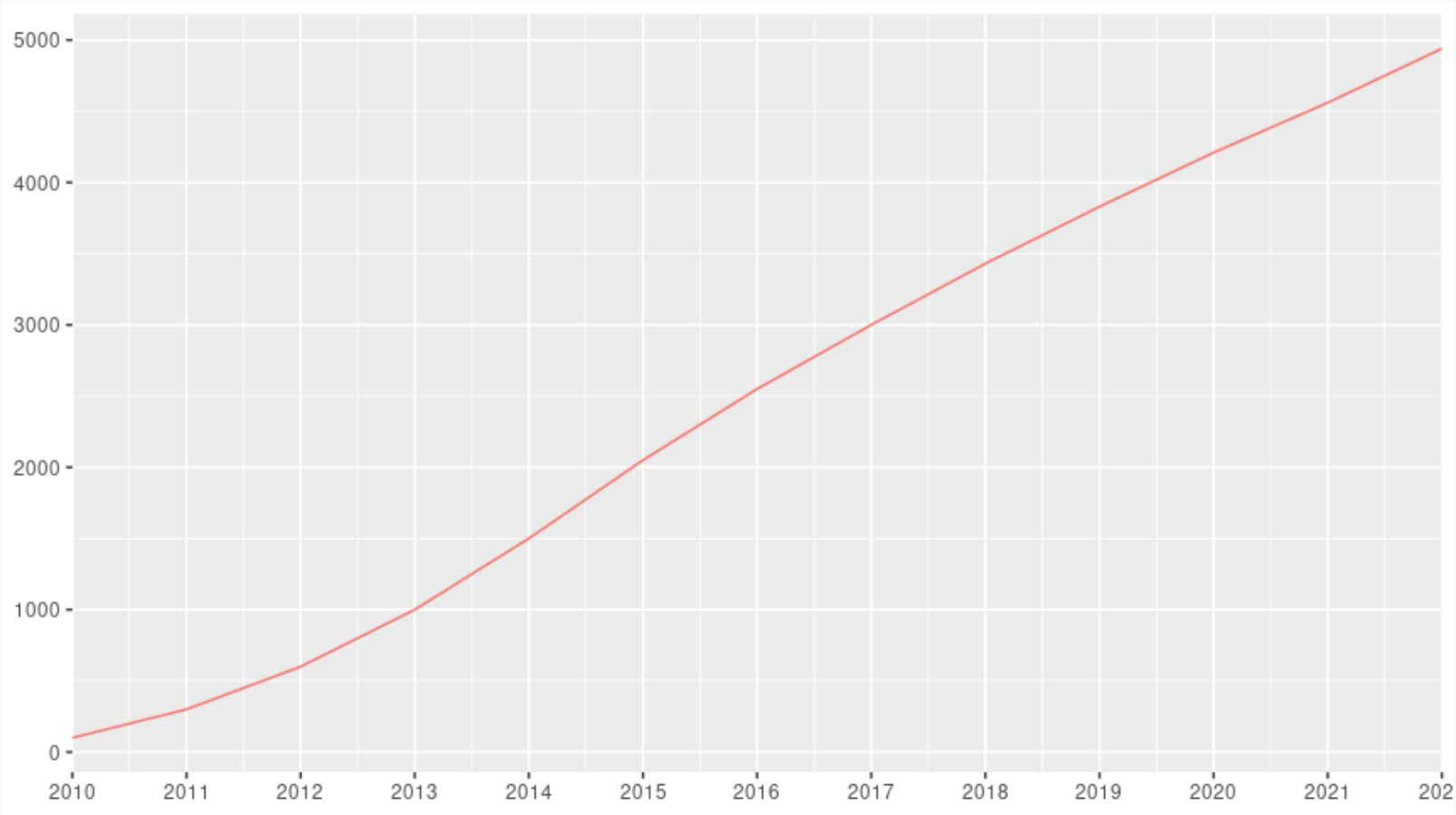


5. Comparez des courbes



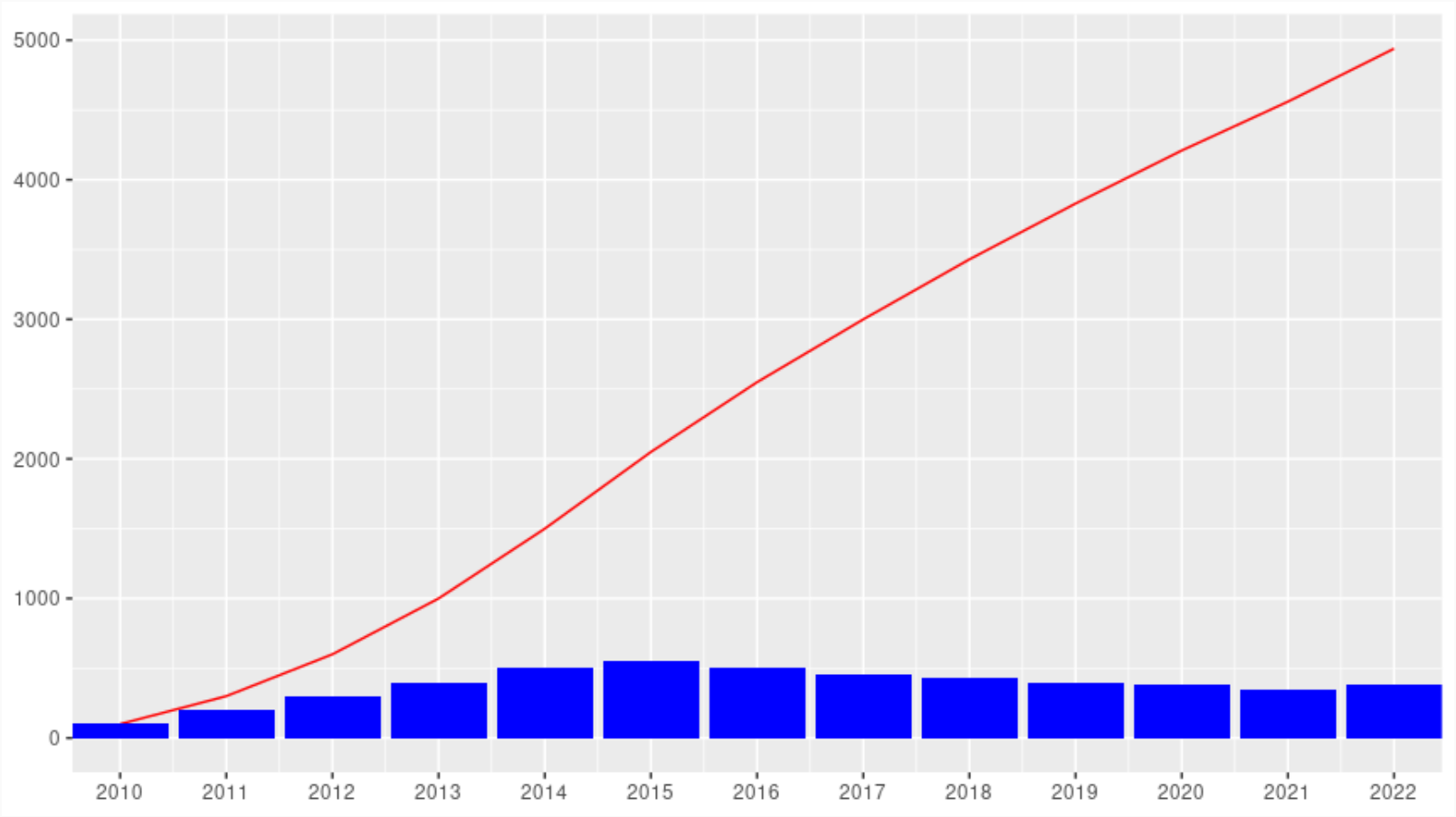
6. Cumulez

Suivez l'exemple de Tim Cook (données simulées)



6. Cumulez

Suivez l'exemple de Tim Cook (données simulées)



7. Utilisez des cartes

Faisons voter les surfaces :

Élections présidentielles aux USA en 2016 (par *counties*)

Source : Lara Trump on Twitter (01/10/2019)

7. Utilisez des cartes

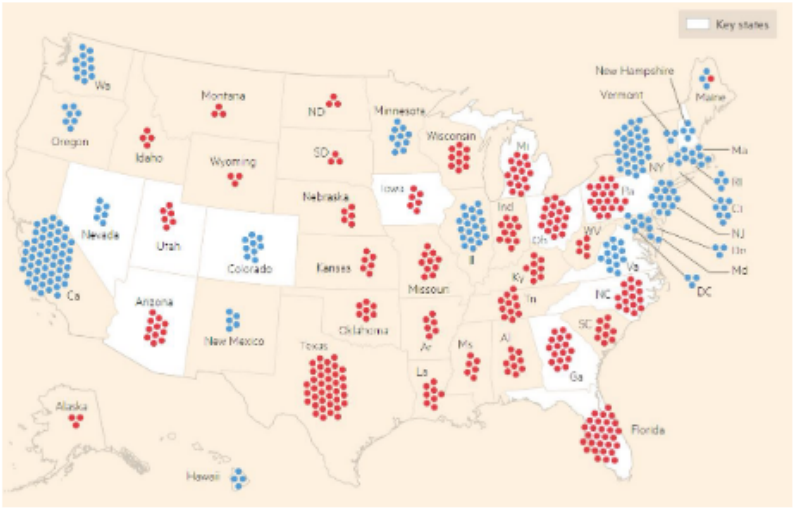
Faisons voter les surfaces :

Élections présidentielles aux USA en 2016 (par *counties*)

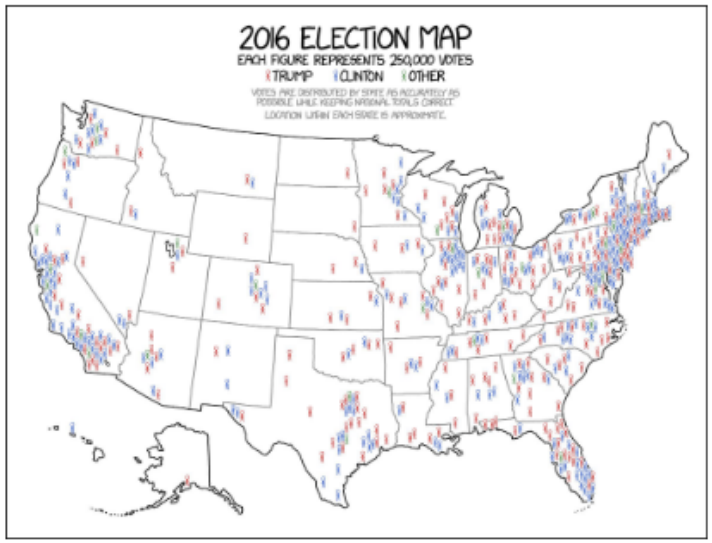
Source : Lara Trump on Twitter (01/10/2019)

7. Utilisez des cartes

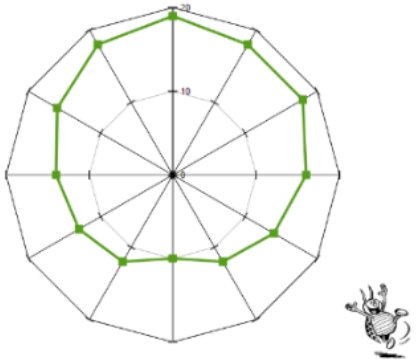
Élections présidentielles aux USA en 2016 (Nombre de grands électeurs)



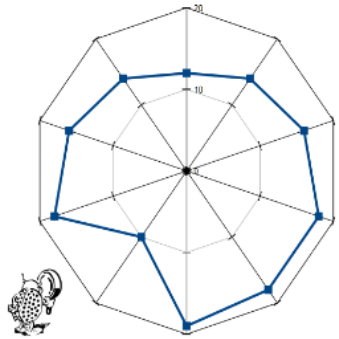
Élections présidentielles aux USA en 2016 (Nombre de votes)



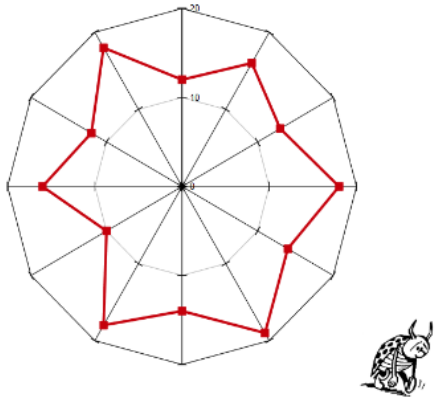
8. Diagrammes en coordonnées polaires



1. profil régulier

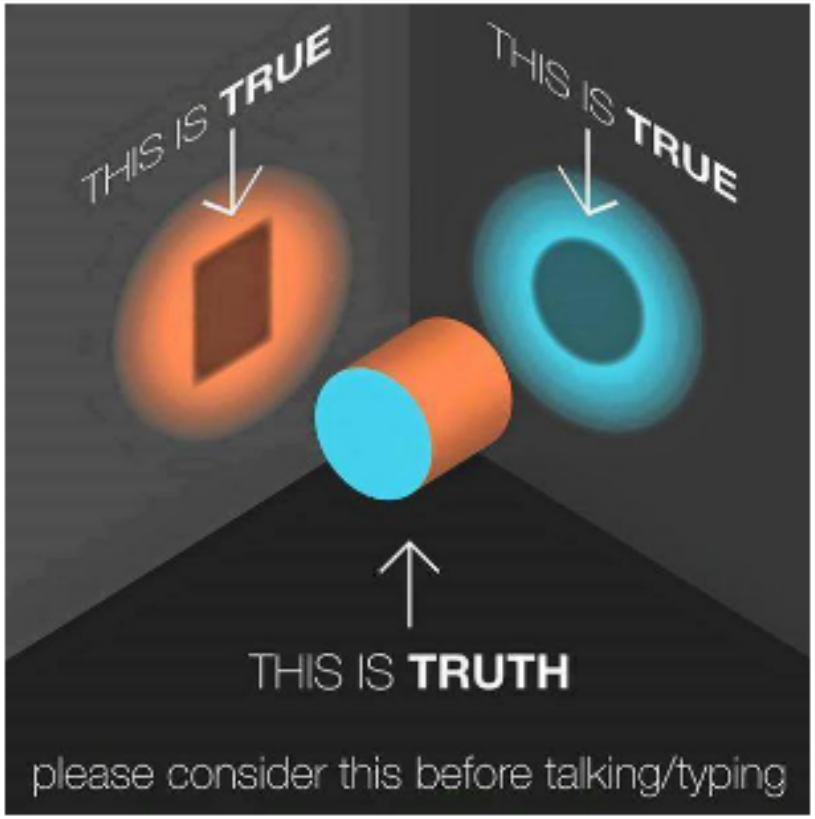


2. profil régulier mais avec une grosse lacune



3. profil irrégulier

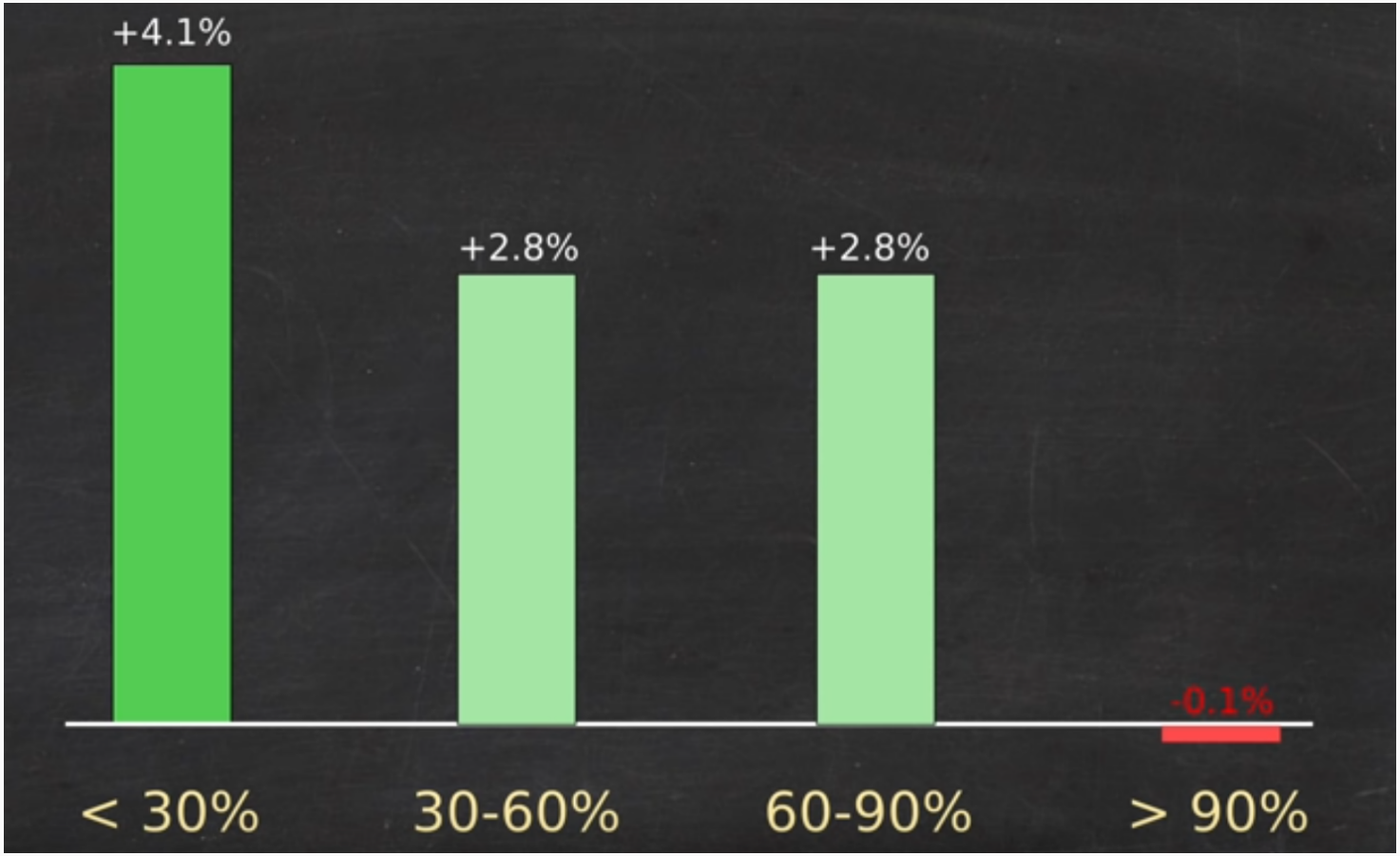
Pour finir


THIS IS TRUE
THIS IS TRUE
THIS IS TRUTH
please consider this before talking/typing

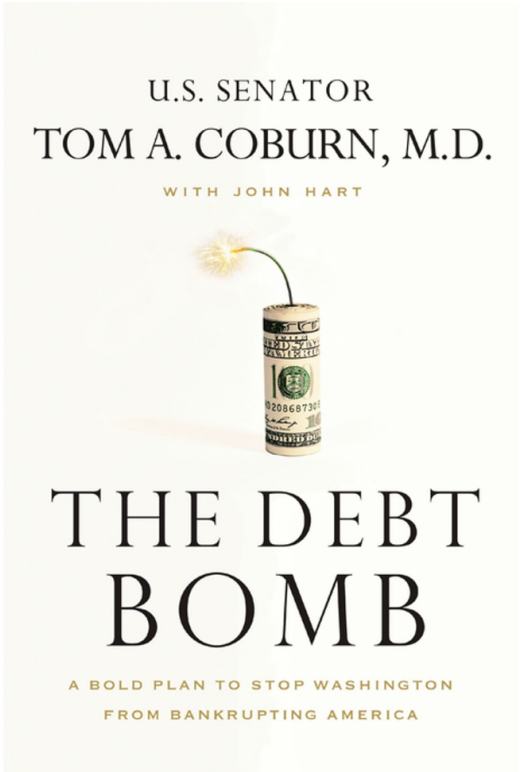
La dette publique : une catastrophe pour la croissance ?

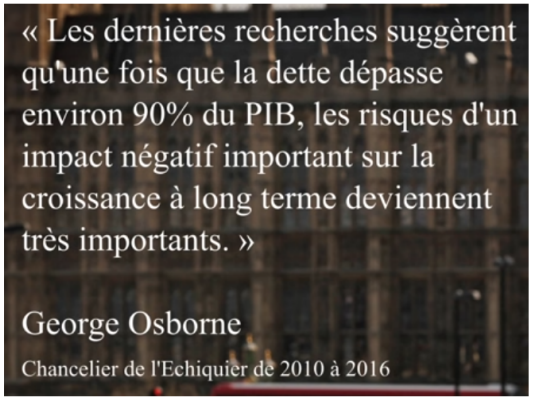
La première étude

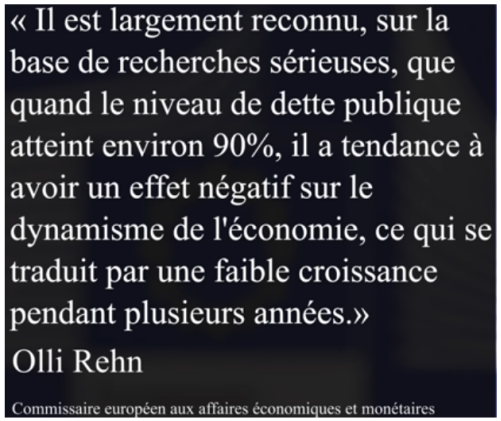
Point de départ : étude (publiée dans « The American Economic Review ») de Reinhart et Rogoff (Harvard) en 2010 sur le lien entre endettement et croissance :



Les conséquences dans le débat public

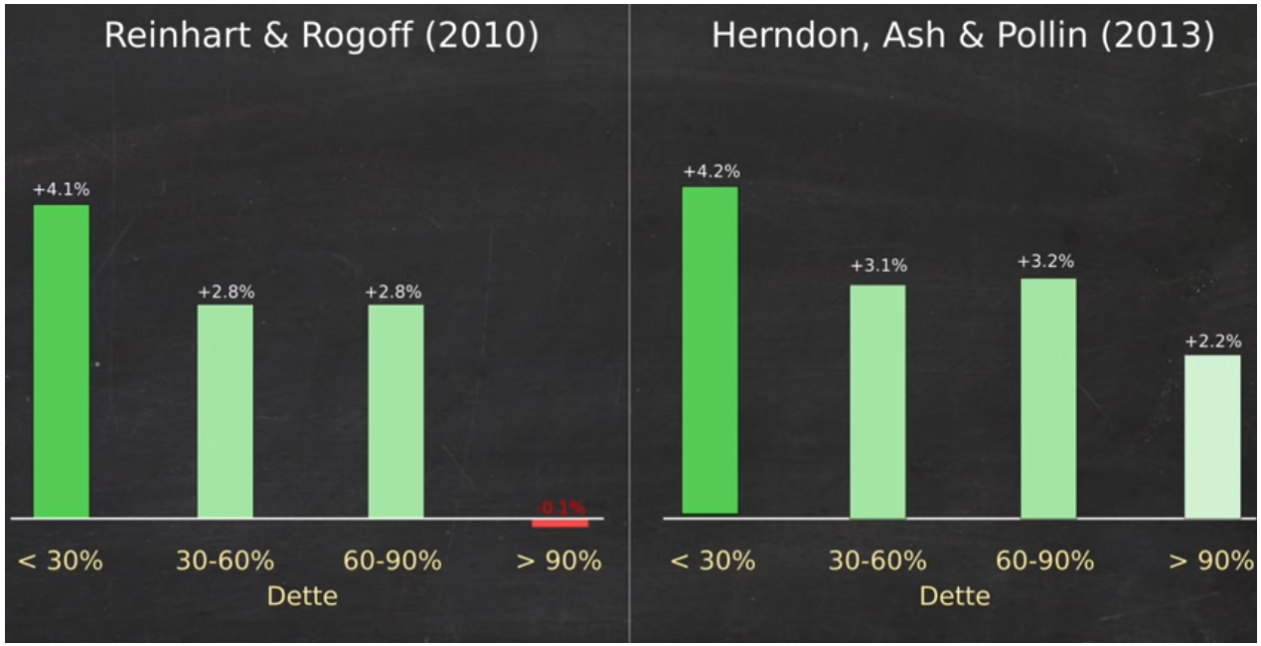






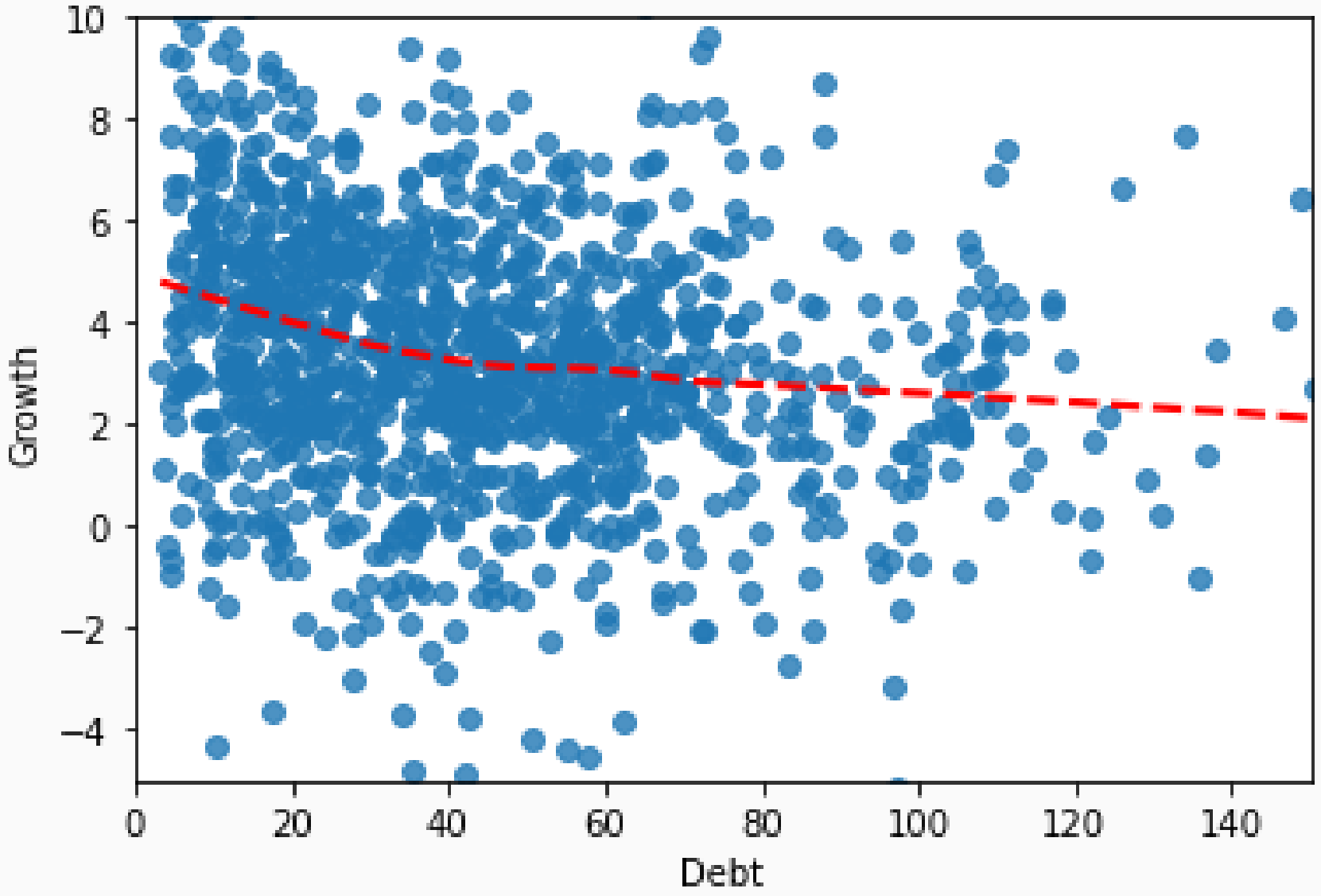
Les résultats corrigés

- 2013 : un étudiant en économie (Thomas Herdnon) tente de reproduire les résultats de l'étude dans le cadre d'un devoir ;
- Trois problèmes sont révélés : pondération étrange des moyennes, exclusion de données sans justification et faute de frappe dans un tableur Excel ...



- Mais le problème est plus profond.

En regardant de plus près



Rendons à César ...

- Chaine Youtube « science étonnante » :

  `https://www.youtube.com/c/ScienceEtonnante`

- Chaine Youtube « la biologie fait des vidéos » :

  `https://www.youtube.com/c/LaBiologiefaitdesvidÃľos`

- Blog de Julien Gossa :

  `https://blog.educpros.fr/julien-gossa/`

- Blog « Data Visualisation » :

  `http://data.visualisation.free.fr/`

- Collectif « Le Cortecs » :

  `https://cortecs.org/`