STATISTICS 102

TODOs

PROBABILITÉS VS STATISTIQUES

Les **probabilités** permettent de représenter un état théorique des choses.

Les **statistiques** utilisent des données (souvent une grande quantité), expérimentales (ou simulées, dans le cadre d'un cours), et permettent de comprendre, d'**estimer les valeurs théoriques**.

ÉCHANTILLON STATISTIQUE

(X1,...,Xn) est appelé **échantillon** si les variables aléatoires X1,...,Xn sont **indépendantes** et suivent la même loi. On dit qu'elles sont **i.i.d.** : **indépendantes et identiquement distribuées.**

Exemples:

- X1...Xn sont n lancers de pile ou face. Ils sont indépendants, et ont tous la même loi de probabilité.
- X1...Xn représentent le QI de n personnes. Ces personnes sont indépendantes et leur QI suit la même distribution (assimilée à loi normale).

Pause

Fini les définitions!
Place à l’estimation

ESTIMATION DE LA MOYENNE

L'espérance de la loi suivie par un échantillon i.i.d. est **estimée** par la **moyenne empirique**:

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Exemples:

- X1…X1000 sont 1000 réponses à la question "êtes-vous satisfait du président" (Oui/Non) : la **côte de popularité** (e.g. 37%) est ***estimée*** en prenant la moyenne empirique des 1000.
- X1…X200 sont le QI de 200 personne. Leur moyenne **estime** la moyenne théorique (l'espérance) du QI.

LA LOI DES GRANDS NOMBRES

Si (X1, ..., Xn) est un échantillon suivant une loi de moyenne μ et de variance σ^2, alors, plus n est grand, plus leur moyenne **empirique** $\overline{X}$ **s'approche** de leur moyenne **théorique** E(X) :

$$\overline{X} \longrightarrow_{n \to \infty} E(X)$$

EXPÉRIENCE

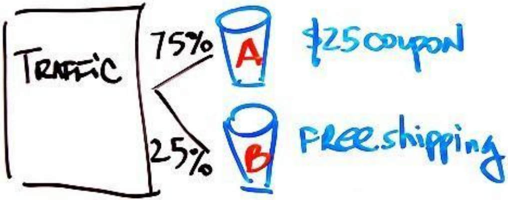
Sur votre téléphone, allez sur internet et cherchez “Pile ou Face”, vous devriez tomber sur une simulation. Faites-le 5 fois et retenez le nombre de fois que vous obtenez “Face” (heads: la tête)

Combien de “Face” sur vos 5 lancers?

(je noterai les résultats et calculerai simplement la moyenne générale au tableau)

Une Application réaliste

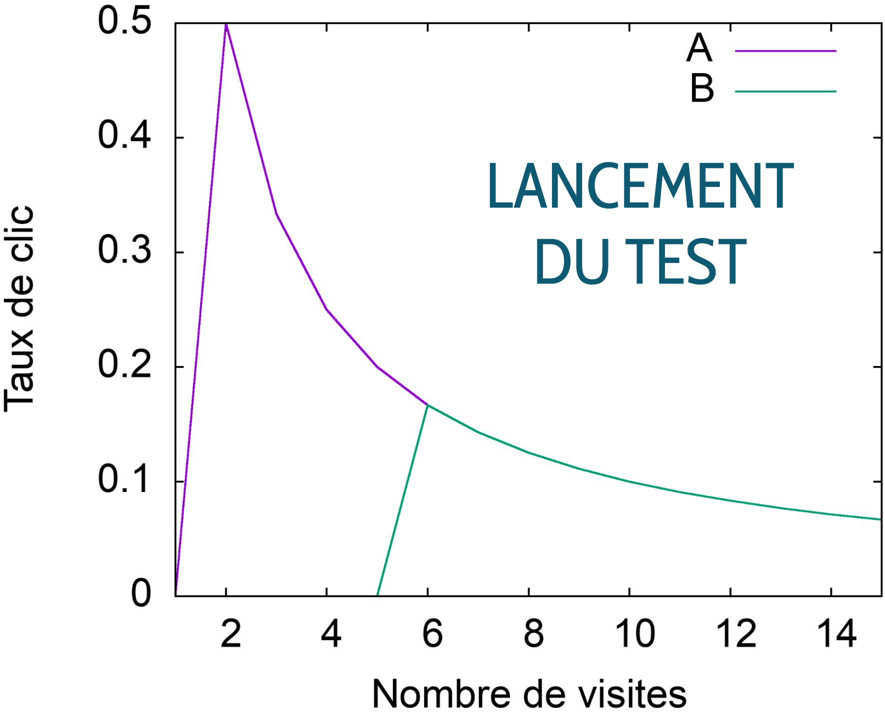
A/B TESTING

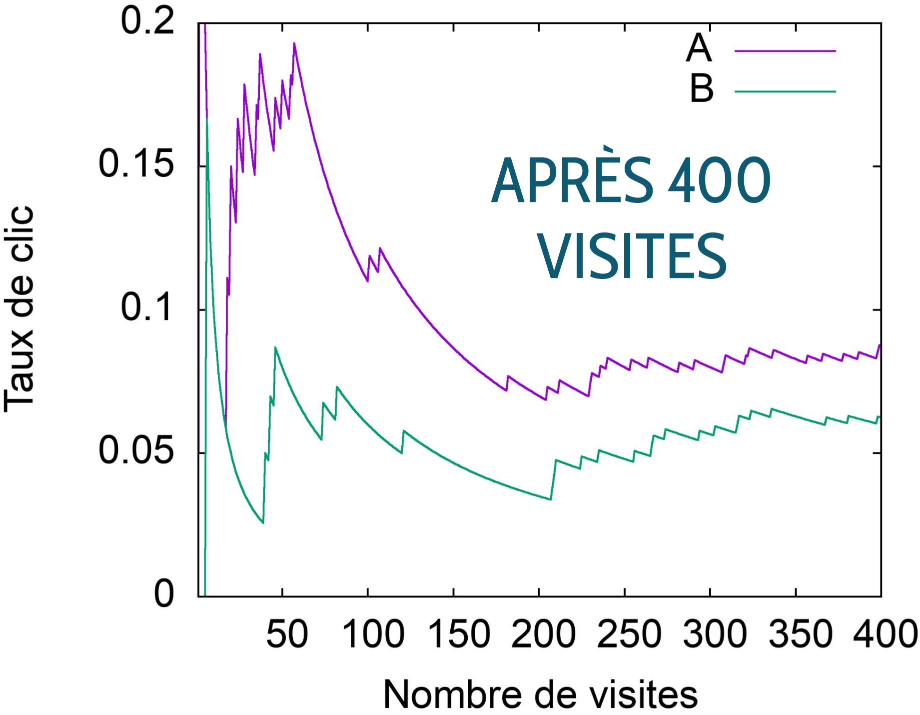


Source: http://www.littleblackdogsocialmedia.com

Source : https://vwo.com

Source : http://www.wordstream.com/


0.5
0.4
0.3
0.2
0.1
0
Taux de clic
A
B
LANCEMENT
DU TEST
2
4
6
8
10
12
14
Nombre de visites


APRÈS 400 VISITES
A
B
Taux de clic
Nombre de visites
0.2
0.15
0.1
0.05
0
50
100
150
200
250
300
350
400

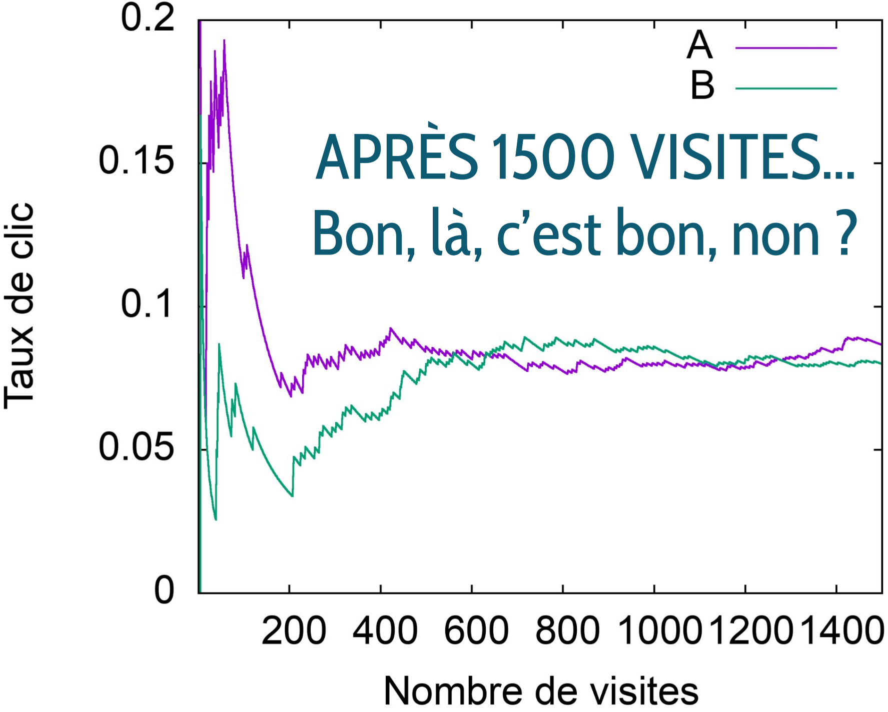
QUAND PRENDRE UNE DÉCISION?

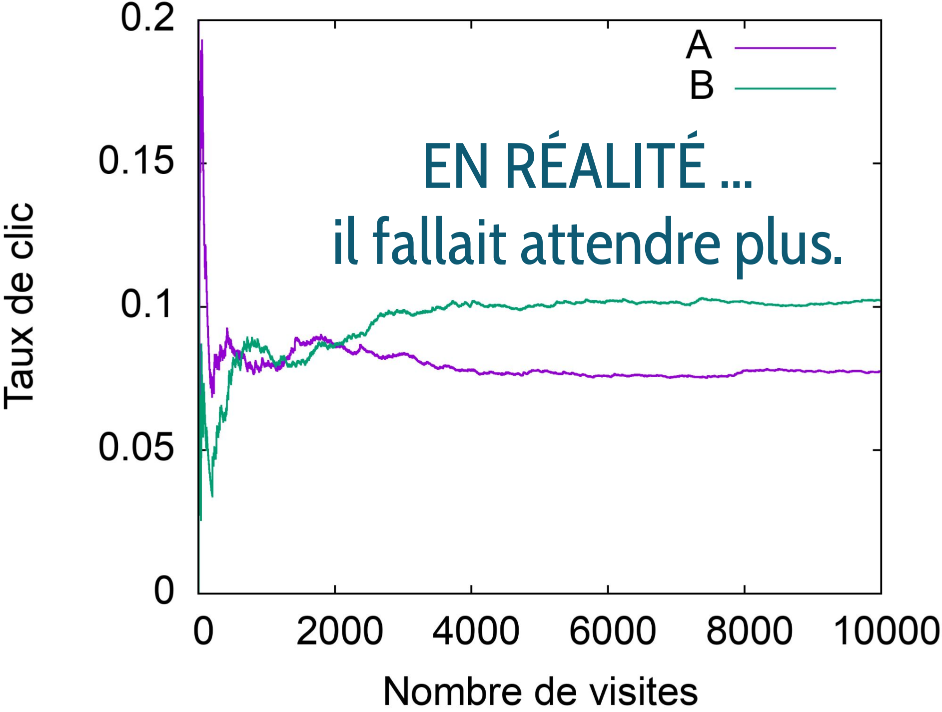
Nous savons déjà (loi des grands nombres) que la moyenne empirique ***tend*** vers la moyenne théorique.   Mais à partir de **quand** ceci est-il **vraiment** applicable ?

On ne peut jamais être sûr (à 100%) du résultat.

Intuitivement, plus l'échantillon est grand et plus la différence entre les deux courbes est forte, plus le résultat est fiable.

On peut en fait **quantifier** la fiabilité du résultat.


APRÈS 1500 VISITES…
Bon, là, c’est bon, non ?
A
B
Taux de clic
0.2
0.15
0.1
0.05
0
200
400
600
800
1000
1200
1400
Nombre de visites


EN RÉALITÉ …
il fallait attendre plus.
A
B
Taux de clic
Nombre de visites
0.2
0.15
0.1
0.05
0
0
2000
4000
6000
8000
10000

QUAND PRENDRE UNE DÉCISION?

Intuitivement, plus l'échantillon est grand et plus la différence entre les deux courbes est forte, plus le résultat est fiable…

MIEUX:

On peut **quantifier** la fiabilité du résultat.

ESTIMATION DE LA VARIANCE

Si (X1, ..., Xn) est un échantillon, on estime sa variance S^2 par la **variance empirique**:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Remarque: La racine S (ou σ) de la variance empirique S^2 est l'**écart-type empirique**. Il estime l'écart moyen à la moyenne observée.

EXERCICE

Calculer la moyenne, la variance empirique et l'écart-type empirique de l'échantillon suivant:

2, 3, 5, 2, 1

EXERCICE

Calculer la moyenne, la variance empirique et l'écart-type empirique de l'échantillon suivant:

2, 3, 5, 2, 1

moyenne :

$$\overline{x} = \frac{2+3+5+2+1}{5} = \frac{13}{5} = 2.6$$

EXERCICE

Calculer la moyenne, la variance empirique et l'écart-type empirique de l'échantillon suivant:

2, 3, 5, 2, 1

moyenne : $$\overline{x} = \frac{2+3+5+2+1}{5} = \frac{13}{5} = 2.6$$

variance :

$$s^2 = \frac{(2-2.6)^2 + (3-2.6)^2 + (5-2.6)^2 + (2-2.6)^2 + (1-2.6)^2}{5-1}$$
$$= \frac{0.6^2 + 0.4^2 + 2.4^2 + 0.6^2 + 1.6^2}{4}$$
$$= \frac{9.2}{4}$$
$$= 2.3$$

EXERCICE

Calculer la moyenne, la variance empirique et l'écart-type empirique de l'échantillon suivant:

2, 3, 5, 2, 1

moyenne : $\overline{x} = \frac{2+3+5+2+1}{5} = \frac{13}{5} = 2.6$

variance : $s^2 = 2.3$

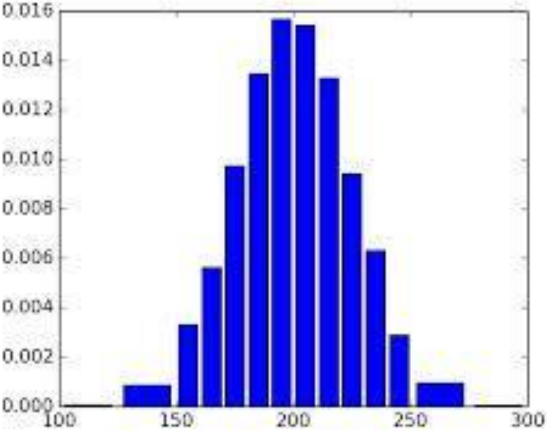
écart-type : $s = \sqrt{2.3} = 1.516$

RETOUR SUR LA LOI NORMALE

Histoire de la loi normale :

- Loi Binomiale: N Lancers Pile/Face (de Moivre, 1756)
- Loi Normale (Laplace, 1781)
- Loi des erreurs (Gauss, 1802)

source : matplotlib.org



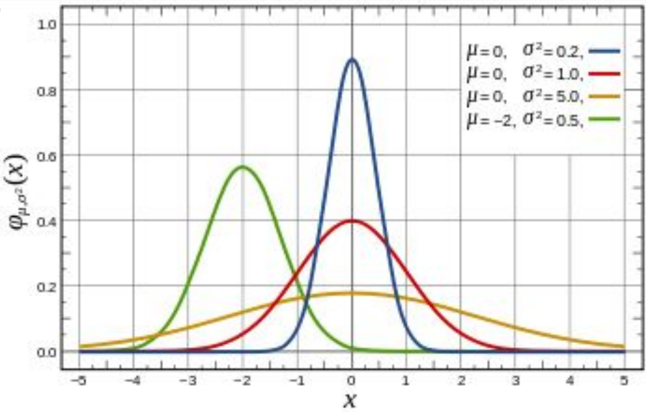
- L'homme moyen (Quételet, 1846)
- La Plance de Galton, L'eugénisme (1889)

RETOUR SUR LA LOI NORMALE (2)

Si X suit la loi normale $\mathcal{N}(\mu, \sigma^2)$ (ou $\mathcal{N}(\mu, \sigma)$)

- X peut prendre toute valeur entre $-\infty$ et $+\infty$.
- La courbe est symétrique.
- **E(X)** = μ (courbe centrée en μ)
- Écart-type = σ
- **V(X)** = σ²
- *Comme toute loi de proba: Aire sous la courbe = 1*

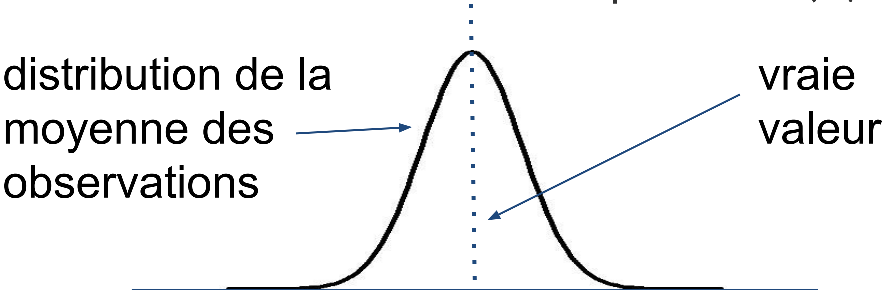
source : wikipedia



THÉORÈME DE LA LIMITE CENTRALE (0)

Loi des grands nombres : si on tire N fois une variable X, la **moyenne** (empirique) $\overline{X}$ tend vers l'espérance (théorique) E(X) quand N $\rightarrow \infty$.

Théorème de la limite centrale : cette moyenne $\overline{X}$ est une V.A. dont la distribution tend vers une **loi normale** centrée sur l'espérance E(X).

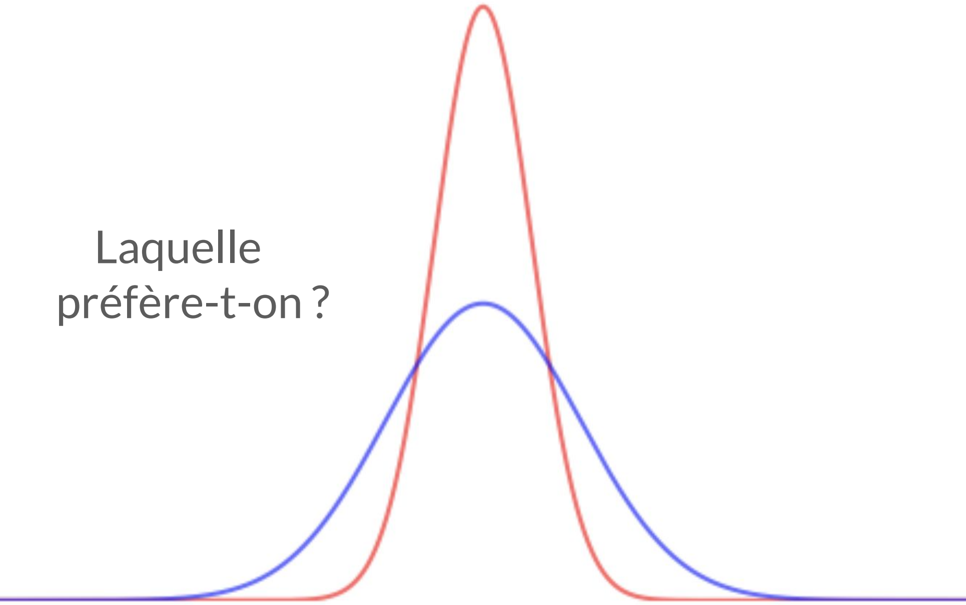


THÉORÈME DE LA LIMITE CENTRALE (1)

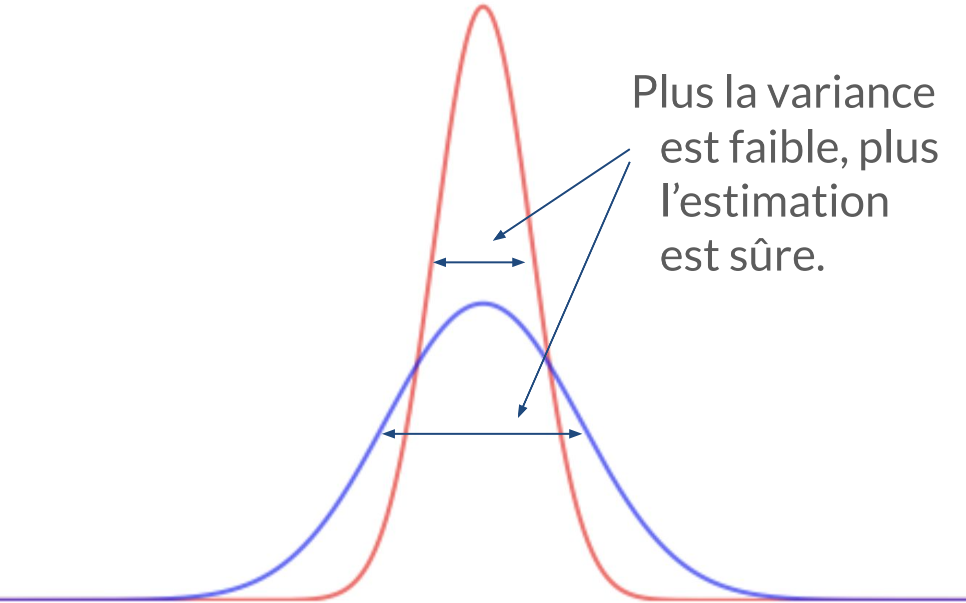
Reprise des chiffres de l'Expérience "Pile ou Face".

Graphe au tableau.

THÉORÈME DE LA LIMITE CENTRALE (2)



THÉORÈME DE LA LIMITE CENTRALE (3)



THÉORÈME DE LA LIMITE CENTRALE (4)

Théorème de la limite centrale : La moyenne empirique d'un échantillon d'observations i.i.d. (indépendantes et identiquement distribuées) s'approche d'une loi normale de paramètres:

moyenne **μ** = moyenne théorique
variance σ^2 = variance théorique / **nombre d'échantillons**

Conclusion : plus vous avez d'échantillons, plus vous pouvez affirmer que vous avez une **bonne** estimation de la vraie moyenne.

THÉORÈME DE LA LIMITE CENTRALE (5)

Un des plus grands résultats statistiques.

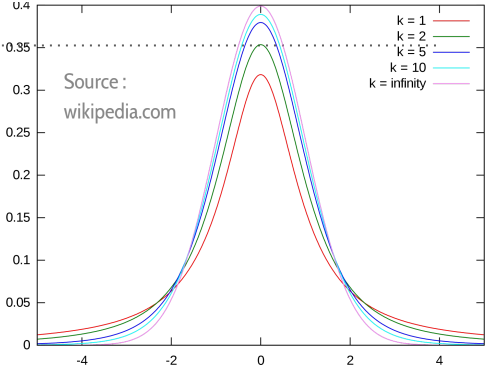
Si (X1, ..., Xn) est un échantillon iid suivant une loi de moyenne μ et de variance σ^2, alors, si n tend vers ∞, leur **moyenne empirique tend vers la loi normale** $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ $\overline{X} \underset{n \to +\infty}{\longrightarrow} \mathcal{N}(\mu, \frac{\sigma^2}{n})$

Ou encore: $$\boxed{\sqrt{n}\,\frac{\overline{X}-\mu}{\sigma} \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1)}$$

Remarque: σ^2/n est la fiabilité de l'estimation: Plus n est grand, plus il est probable que $\overline{X} \approx \mu$

ESTIMATION PAR LA LOI DE STUDENT

Lorsque la variance de l'échantillon n'est pas connue, la loi normale est remplacée par la loi de **Student** de paramètre n − 1.



Lorsque l'échantillon (n) est suffisamment grand la loi de Student s'apparente à une loi normale.

Pour simplifier, on admet:

$$\sqrt{n}\,\frac{\overline{X}-\mu}{s} \xrightarrow[n\to\infty]{} \mathcal{N}(0,1)$$


FLEX TAPE
STRONG RUBBERIZED WATERPROOF TAPE
PHIL SWIFT


Unknown distribution
Data Scientist
Gaussian

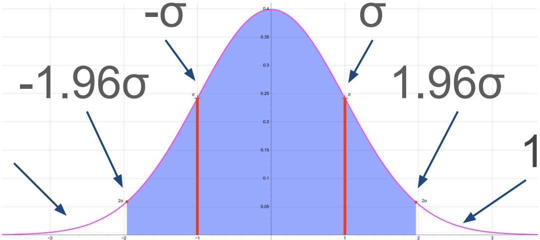
INTERVALLE DE CONFIANCE

Dans le cas d'un échantillon (X1, … , Xn) de moyenne empirique $\overline{X}$, la moyenne théorique μ est avec une certitude de 95% dans l'intervalle:

$$\left[\overline{X} - 1.96\frac{s}{\sqrt{n}},\ \overline{X} + 1.96\frac{s}{\sqrt{n}}\right]$$

c.à.d: $$P\left(-1.96 < \sqrt{n}\,\frac{\overline{X}-\mu}{s} < 1.96\right) = 0.95$$

Aire à gauche de -1.96: 2.5% = 0.025



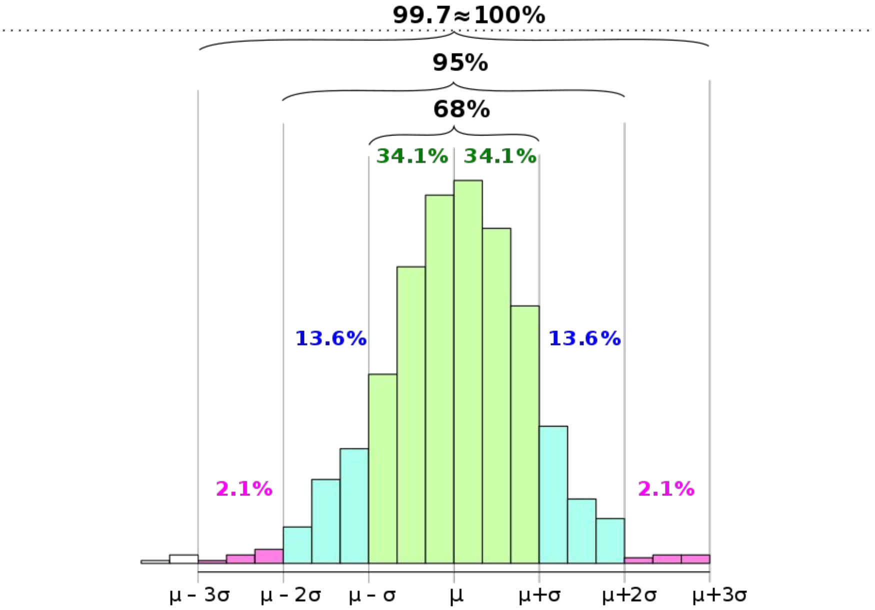
Aire à droite de 1.96: 2.5% = 0.025

INTERVALLE DE CONFIANCE

Pause / Autre exemple

au tableau

LOI 68-95-99.7



LOI 68-95-99.7

Range	Proba. Inside range	Proba. outside range	Freq. for daily event
μ ± σ	**0.68**2689492	1 / 3	Twice a week
μ ± 2σ	**0.95**4499736	1 / 22	Every 3 weeks
μ ± 3σ	**0.997**300203	1 / 370	Yearly
μ ± 4σ	0.999936657	≈ 1 / 15800	Every 43 years
μ ± 5σ	0.999999426	≈ 1 / 1.74M	Every 4776 years
μ ± 6σ	0.999999998	≈ 1 / 507M	Every 1.38M years
μ ± 7σ	0.999999999997	≈ 1 / 391G	Every 1.07B years

LES TESTS STATISTIQUES

Vous voulez prendre une décision et vous voulez connaître la probabilité que cette décision **soit une erreur**. Par exemple, si le taux de clics est de 4% ou plus, on décidera de dépenser 1 million de plus dans la pub.

Vous observez que sur une semaine, le taux de clic moyen est de 5.1%, sur $n = 10000$ visites.

LES TESTS STATISTIQUES

Vous voulez prendre une décision et vous voulez connaître la probabilité que cette décision **soit une erreur**. Par exemple, si le taux de clics est de 4% ou plus, on décidera de dépenser 1 million de plus dans la pub.

Vous observez que sur une semaine, le taux de clic moyen est de 5.1%, sur n = 10000 visites.

La moyenne empirique est **plus grande** que la valeur seuil de 0.04: on est bon! (si elle était plus petite, sans faire de calcul, à priori c'était foutu).

LES TESTS STATISTIQUES

Rappel: n = 10000 visites, taux de clic 5.1%.
On veut savoir avec quelle certitude $\mu \geq 0.04$. Ou encore, la probabilité qu'on se trompe, qui est la probabilité que $\mu < 0.04$.

Il est impossible de calculer cette probabilité directement. Toutefois on peut en calculer une autre, similaire, qui nous servira de **borne:**
La probabilité que, étant dans le cas défavorable ($\mu < 0.04$) le plus "limite" possible (donc $\mu = 0.04$ à la limite), on ait pu tout de même observer 0.051 ou plus.

LES TESTS STATISTIQUES

On veut donc estimer:

P (taux ≥ 0.051 sur n=10000 visites | μ = 0.04)

Rappel: $$\sqrt{10000}\,\frac{0.051-\mu}{0.22} \rightsquigarrow \mathcal{N}(0,1)$$

En remplaçant μ par 0.04, on obtient la valeur 5, qui est un nombre très excentré dans la distribution: la probabilité d'obtenir 5 **ou plus grand** en tirant une loi normale N(0,1) est très faible (1 sur 3.4million).

C'est, à peu près, une borne *supérieure* de la probabilité que vous vous trompez.

LES TESTS STATISTIQUES

Supposons qu'on ait observé un taux de clic de **4.11%** sur n=10000. En reprenant la formule:

Rappel: $$\sqrt{10000}\,\frac{0.0411-\mu}{0.22} \rightsquigarrow \mathcal{N}(0,1)$$

En remplaçant µ par 0.04, on obtient cette fois 0.5. La probabilité d'obtenir 0.5 *ou plus grand* en tirant une loi normale **N**(0,1) est d'environ 1/3. Autrement dit, P(µ < 0.04) est à priori < 1/3, mais ce n'est pas assez faible pour être sûr: vous ne concluez donc **pas** que l'experience vous donne assez d'information.

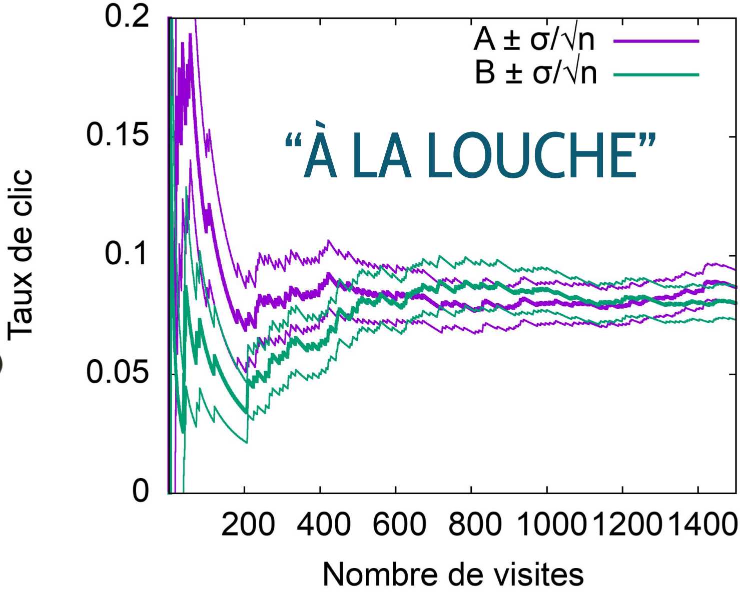
LES TESTS STATISTIQUES (2)

Pour généraliser:

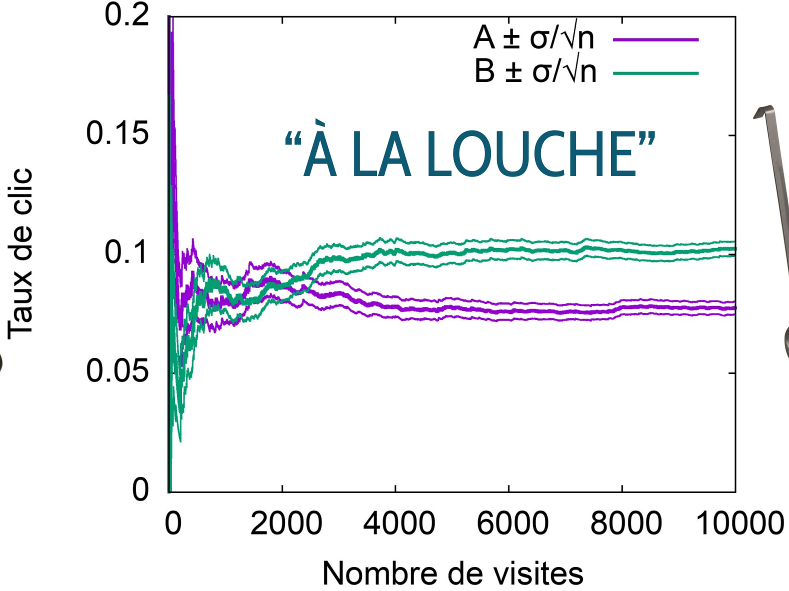
J'observe un échantillon de taille n et de moyenne empirique $\overline{X}$ et j'en conclus que la moyenne théorique μ est **≥** μ_0 (une constante).

Avec quelle probabilité ai-je tort ?

- Le calcul suivant ne marche que si $\mu_0 < \overline{X}$
- Je calcule t = $\sqrt{n}\,\frac{\overline{X}-\mu_0}{s}$
- Je calcule P(X ≥ t) pour X v.a. de loi $\mathcal{N}(0,1)$
- La valeur obtenue est la **p-value**. C'est une borne *supérieure* de la probabilité que je me trompe $P(\mu < \mu_0)$.



Calcul d’écart-type (empirique!) σ, et on ajoute les courbes à ±σ/√n (68%), ±2σ/√n (95%),...



… On aurait su qu’il fallait continuer !

“À LA LOUCHE”

Loi binomiale (N tirages de Bernoulli i.i.d) de paramètre p, puis on fait la moyenne.

Ici X est la **moyenne** des N tirages.

- $E(X) = p$
- $V(X) = pq = p(1-p)/N$ donc **$\sigma(X) = \sqrt{(pq/N)}$**

  (si le temps le permet, re-calcul au tableau)

"À LA LOUCHE"

Loi binomiale: X moyenne de N oui/non, P(oui)=p

$E(X) = p$, $\sigma(X) = \sqrt{pq/N}$

Présidentielles:

- Un sondage donne 60% d'intentions de votes pour A. Échantillon = 1000 personnes.
  - En supposant i.i.d. (faux!), on obtient: **$E(X) = 60\%$ $\sigma(X) = 1.5\%$**
- Si 5% d'intention de votes: **$E=5\%$ $\sigma=0.7\%$**
- Avec **1%**: **$\sigma=0.3\%$** soit ⅓ de la valeur!

En pratique, ce n'est pas iid: c'est pire!

"À LA LOUCHE"

Loi binomiale: X moyenne de N oui/non, P(oui)=p

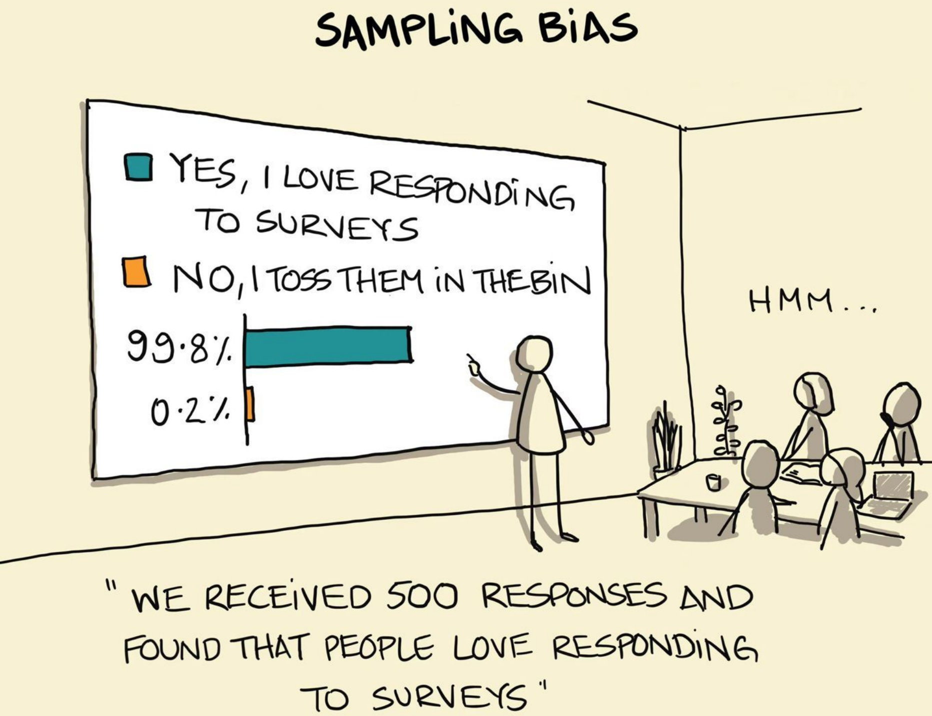
E(X) = p, σ(X) = √(pq/N)

Présidentielles:

- Un sondage donne 60% d'intentions de votes pour A. Échantillon = 1000 personnes.
  - En supposant i.i.d. (faux!), on obtient:
    E(X) = 60% σ(X) = 1.5%
- Si 5% d'intention de votes: **E=5% σ=0.7%**
- Avec **1%**: **σ=0.3%** soit ⅓ de la valeur!

En pratique, ce n'est pas iid: c'est pire!

…. Et tout ça sans compter les **biais** :



Source: sketchplanations.com

CONCLUSION

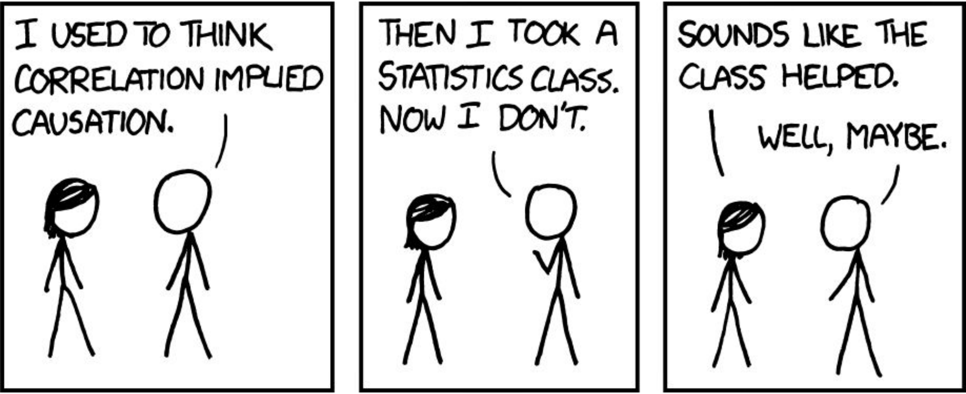
Grâce aux **tests statistiques** on peut:

- vérifier si une impression est justifiée
- donner une conclusion **quantifiée** en termes de risque

C'est une introduction superficielle, il existe de nombreux autres tests!

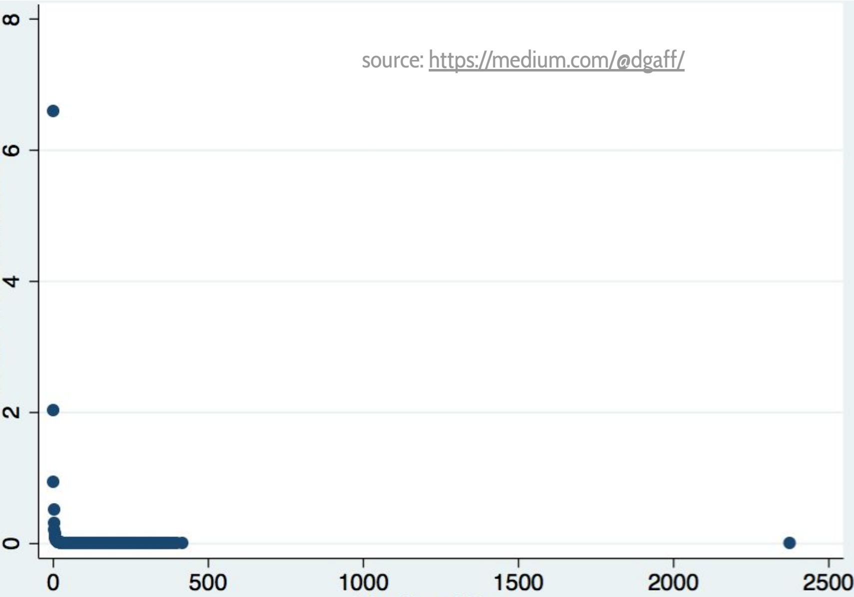
⚠️: difficile de tout bien faire, sans se tromper.

- faire des hypothèses, bien choisir les probas regardées.
- évènements **i.i.d.** pour utiliser les outils liés aux lois normales.



Source: https://xkcd.com/552

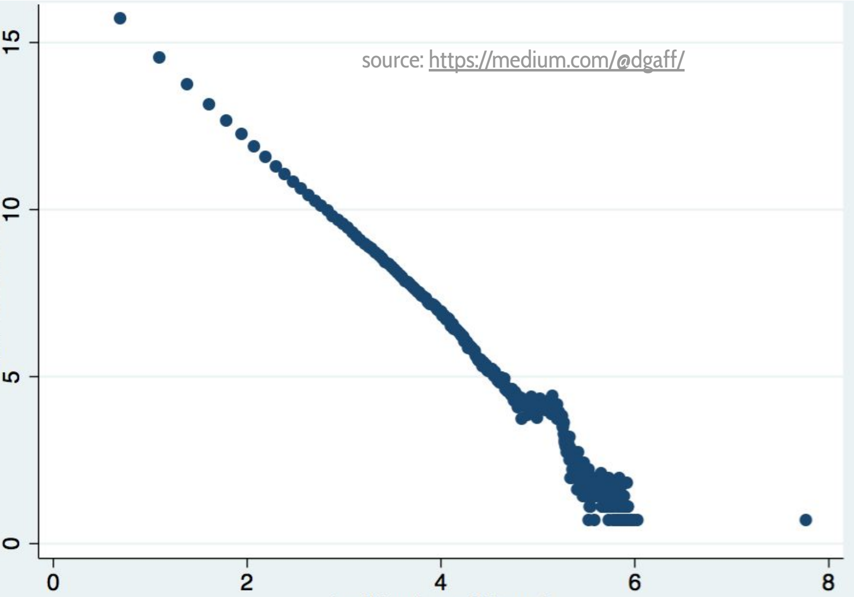
LOIS DE PUISSANCE



Y = Nb de millions d'utilisateurs (faisant ce nb de tweets par jour)

X = Nombre de tweets par jour

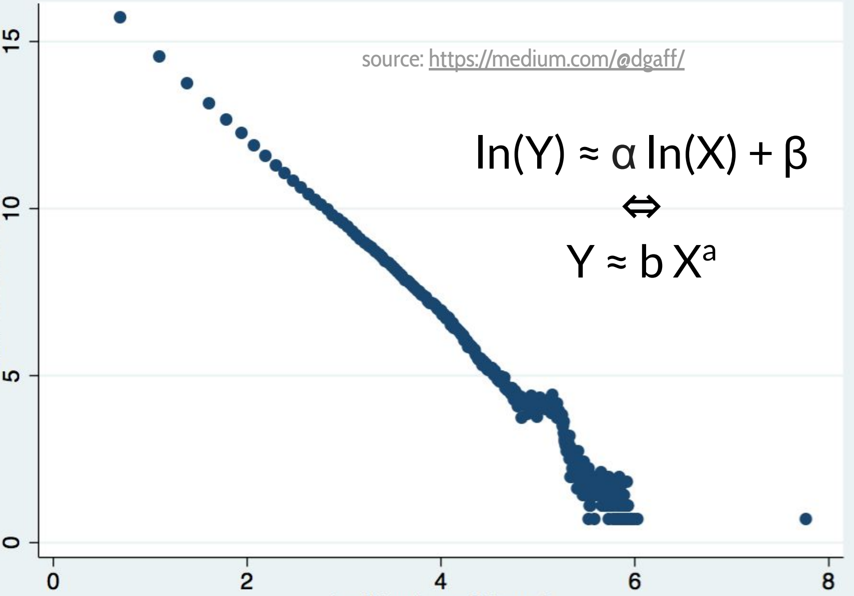
LOIS DE PUISSANCE



ln(Y) = ln(Nombre d'utilisateurs)

ln(X) = ln(Nombre de tweets par jour)

LOIS DE PUISSANCE



$$\ln(Y) \approx \alpha \ln(X) + \beta$$

$$\Leftrightarrow$$

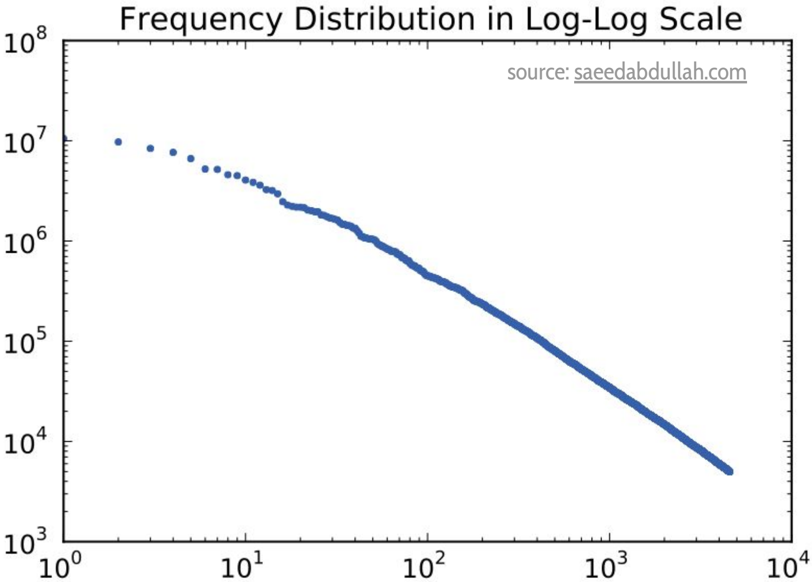
$$Y \approx b\, X^{a}$$

ln(Y = Nombre d'utilisateurs)

ln(X = Nombre de tweets par jour)

LOIS DE PUISSANCE

Y = Nombre d'occurences



X = N^{ème} mot plus fréquent sur Twitter

LOIS DE PUISSANCE, ET AUTRES …

Les lois de puissance apparaissent souvent en pratique (dans la nature aussi!).

La variance peut être **infinie** ! → on **bricole**. Exemple: estimation de la moyenne d'une loi de puissance: impossible "tel quel" si $\alpha < 2$.

Autre loi fréquente: loi **exponentielle.**

Exemple: Il pleut. Δt = temps jusqu'à la prochaine goutte de pluie qui me tombe dessus. Δt est une v.a. qui suit une loi exponentielle. → loi moins problématique (espérance, σ, …).