STATISTICS 101

NOUS PARLERONS DE

Variables aléatoires

Loi de probabilité

Echantillons

Significativité

Tests statistiques

EXEMPLE 1

On cherche à répondre au type de questions suivantes:
Sur 1000 personnes (500 hommes, 500 femmes), on observe que les femmes gagnent en moyenne 2000 euros et les hommes 2100.

Peut-on en déduire que les femmes gagnent moins que les hommes?

Même question si l'on a 200 femmes et 200 hommes.

Même question si l'on a 10 femmes et 10 hommes.

EXEMPLE 2

On parle de **A**/**B** testing.
Sur une page web, il y 2 boutons: un rouge et un bleu.
Sur 1000 personnes l'ayant vu, 23 ont cliqué sur le bouton rouge.

Sur 500 personnes l'ayant vu, 17 ont cliqué sur le bouton bleu.

Peut-on affirmer que le bouton bleu a plus de succès?

EXEMPLE 3

Peut-on affirmer qu'il y a plus de garçons que de filles en informatique?

Peut-on affirmer que l'informatique est la matière où il y a la plus grande différence entre le nombre de garçons et le nombre de filles?

Peut-on affirmer qu'en informatique la plus grande différence entre le nombre de garçons et le nombre de filles est plus grande qu'en histoire?

LA SIGNIFICATIVITÉ

On se pose donc la question de la fiabilité d'un chiffre et de la validité des conclusions.

DÉFINITION

Un résultat est **statistiquement significatif** à 5% si la probabilité qu'on l'observe par hasard est inférieure à 5%.

Intuitivement, cela a un rapport avec:

La taille de l'échantillon

L'hétérogénéité de l'échantillon.

VARIABLES ALÉATOIRES

Définition (pas très mathématique)
Une variable aléatoire (v.a.) est une application définissant l'ensemble des résultats possibles pour une expérience donnée.
Une variable aléatoire est donc la chose que l'on **observe** et qui nous intéresse.

LES VARIABLES DISCRÈTES

Définition

Une variable aléatoire (v.a.) est dite **discrète** lorsqu'elle peut prendre un nombre **dénombrable** de valeurs.

LES VARIABLES DISCRÈTES

Définition

Une variable aléatoire (v.a.) est dite **discrète** lorsqu'elle peut prendre un nombre **dénombrable** de valeurs.

Exemples:

X représente la variable aléatoire liée à l'expérience "pile ou face": les valeurs possibles de X sont 0 et 1.

X représente la variable aléatoire liée à l'expérience "réponse à un sondage": les valeurs possibles de X sont "pas satisfait", "plutôt satisfait", "très satisfait".

X représente la variable aléatoire liée à l'expérience "nombre de personnes dans un supermarché": les valeurs possibles de X sont 0, 1, 2, . . . , +1.

LES VARIABLES CONTINUES

Définition

Une variable aléatoire (v.a.) est dite **continue** lorsqu'elle peut prendre un nombre **indénombrable** de valeurs.

LES VARIABLES CONTINUES

Définition

Une variable aléatoire (v.a.) est dite **continue** lorsqu'elle peut prendre un nombre **indénombrable** de valeurs.

Exemples:

X représente la variable aléatoire liée à l'expérience "heure d'arrivée d'un ami": les valeurs possibles de X se trouvent entre 15h et 16h.

X représente la variable aléatoire liée à l'expérience "poids": les valeurs possibles de X se trouvent entre 0 et +1.

X représente la variable aléatoire liée à l'expérience "Note moyenne en M2": les valeurs possibles de X se trouvent entre 0 et 20.

LOI DE PROBABILITÉ

Définition (pas mathématique du tout!)
Une loi (ou distribution) de probabilité représente le comportement d'une variable aléatoire.

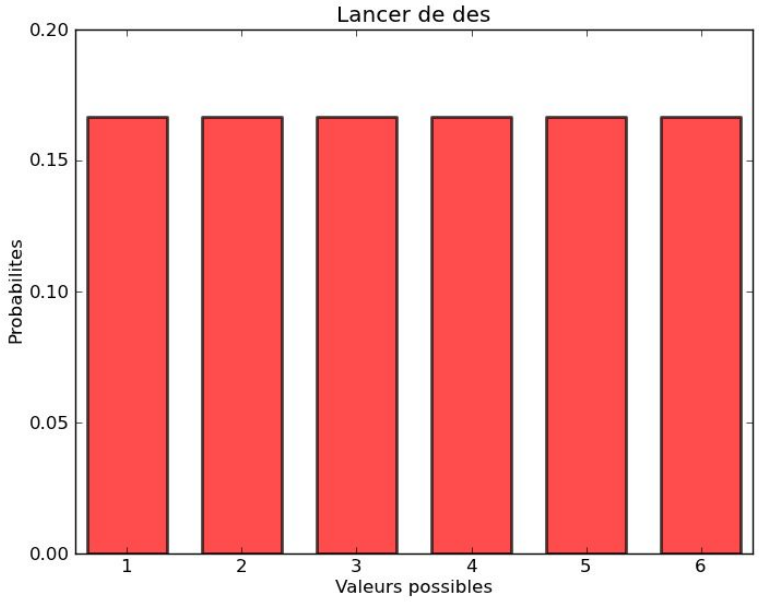
Autrement dit, la loi définit les probabilités qu'une variable aléatoire prenne une valeur ou un ensemble de valeurs. Attention : elle ne dit pas quelle valeur va être prise !

EX 1: LOI UNIFORME DISCRETE

X représente la variable aléatoire liée à l'expérience **"lancer de dé"**: X prend ses valeurs entre 0 et 6.

$P(X = 1) = P(X = 2)$
$= P(X = 3) = P(X = 4)$
$= P(X = 5) = P(X = 6) = 1/6$

La somme des probabilités vaut 1.



Loi loi est dite **uniforme** car toutes les probabilités sont les mêmes.

On parle aussi d'**équiprobabilité**.

EX 2: LOI DE BERNOULLI

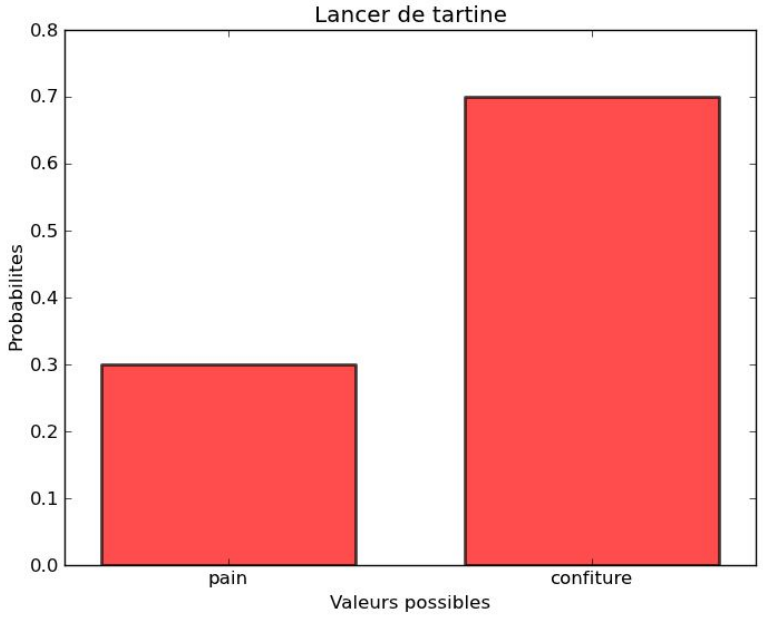
X représente la variable aléatoire liée à l'expérience

"lancer de tartine": X prend les valeurs 0 ou 1.

P(X = côté pain) = 0, 3

P(X = côté confiture) = 0, 7

La somme des probabilités vaut 1.



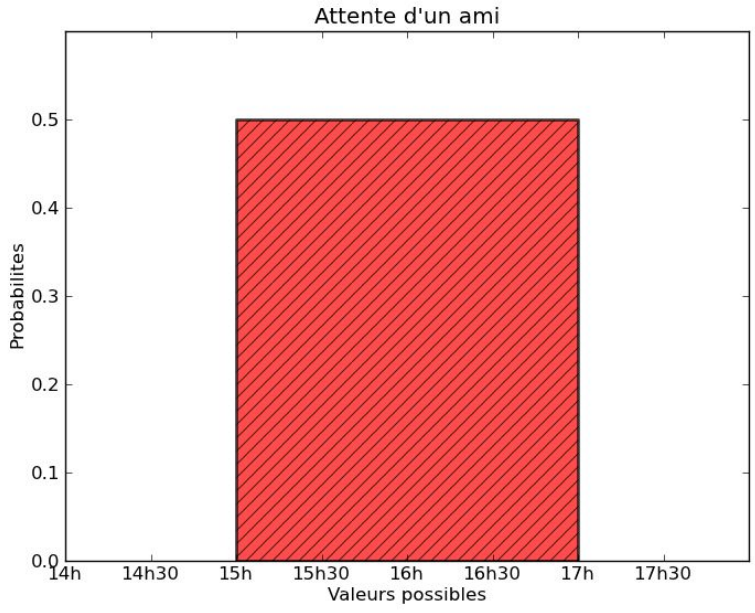
Le **paramètre** de la loi est p.

Ici, p = 0, 7.

Dans le cas d'un pile ou face avec une pièce équilibrée, p = 0, 5.

EX 3: LOI UNIFORME CONTINUE

X représente la variable aléatoire liée à l'expérience **"heure d'arrivée d'un ami"**: X prend ses valeurs entre 15h et 17h.



P(X = 15 : 00) = 0

P(X = 15 : 30) = 0

P(X 2 [15 : 10, 15 : 40]) = 0, 25

P(X > 16 : 15) = 0, 375

La "somme" des probabilités vaut 1. Ici, il s'agit de l'aire sous la courbe.

EX 4: LOI NORMALE

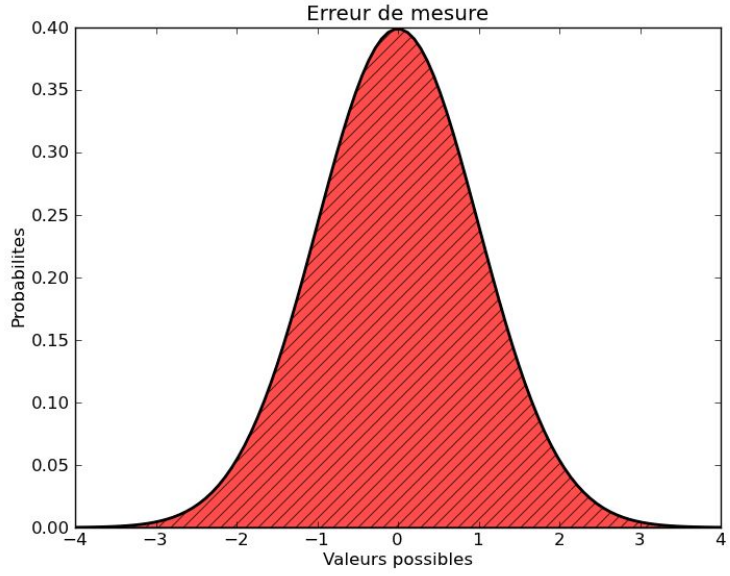
X représente la variable aléatoire liée à l'expérience **"erreur de mesure"**: X prend ses valeurs entre −inf et +inf.

P(X = 0) = 0

P(X = 3) = 0

P(X app [−1.96, 1.96]) = 0, 95

P(X > 2.33) = 0, 01



La "somme" des probabilités vaut 1. Ici, il s'agit de l'aire sous la courbe.

Ici, les deux **paramètres** de la loi sont : moyenne = 0, écart-type = 1. (Nous y reviendrons)

MOMENTS D'UNE V.A.

Définition

Les moments d'une variable aléatoire sont des **indicateurs de tendance** de la variable. Nous ne nous intéresserons qu'à deux d'entre eux : l'**espérance** et la **variance**.

L'**espérance** correspond à la moyenne d'une variable aléatoire. On la note E(X).

La **variance** représente la dispersion d'une variable aléatoire autour de sa moyenne. On la note V(X). Plus la variance est grande, plus la variable est dispersée, hétérogène, imprévisible.

PROBABILITÉS VS STATISTIQUES

Les **probabilités** représentent un état théorique des choses. On ne les connaît pas. Les **statistiques** utilisent des données permettant de comprendre, d'estimer les valeurs théoriques.

ÉCHANTILLON STATISTIQUE

$(X_1,...,X_n)$ est un **échantillon** si les variables aléatoires $X_1,...,X_n$ sont **indépendantes** et suivent la même loi. On dit alors qu'elles sont **indépendantes et identiquement distribuées** (i.i.d.).

Exemples:

$X_1...X_n$ sont n lancers de pile ou face. Ils sont indépendants et ont tous la même loi de probabilité.

$X_1...X_n$ représentent le QI de n personnes. Ces personnes sont indépendantes et leur QI suit la même distribution (assimilée à une loi normale).

ESTIMATION DE LA MOYENNE

L'espérance de la loi suivie par un échantillon 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:

$X_1 \ldots X_{1000}$ sont 1000 réponses à la question "êtes-vous satisfait de l'action présidentielle?": la **côte de popularité** est estimée sur la réponse de ces 1000 personnes en prenant la moyenne.

$X_1 \ldots X_{200}$ représentent le QI pour 200 personne : leur moyenne **estime** la moyenne théorique (l'espérance).

LA LOI DES GRANDS NOMBRES

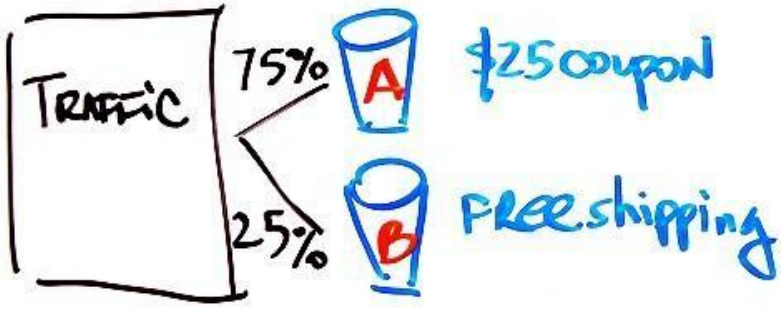
http://bitly.com/2gsd1S3

LA LOI DES GRANDS NOMBRES

Si (X_1, ..., X_n) 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 (m)** :

$$\overline{X} \longrightarrow_{n \to \infty} m$$

A/B TESTING

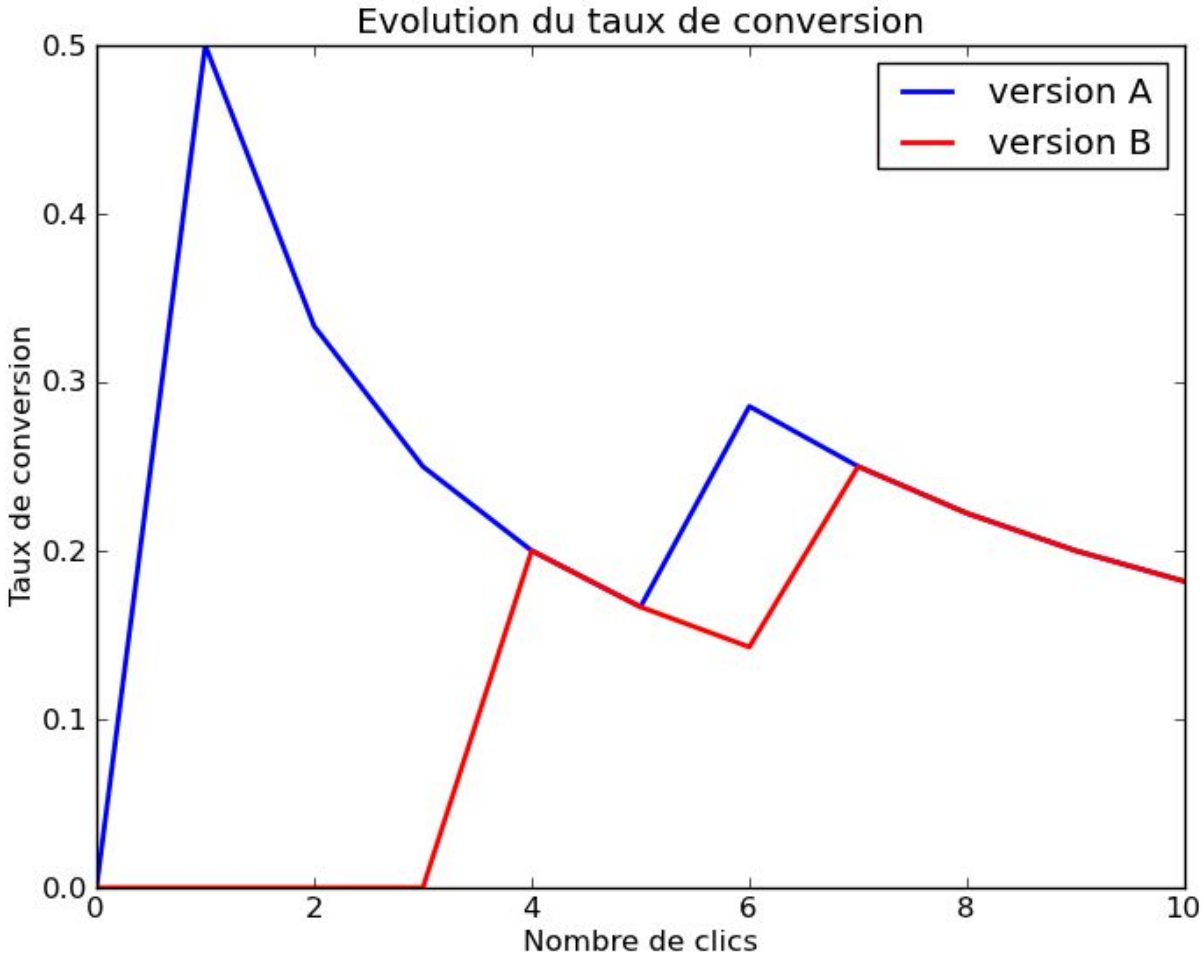


Source:
http://www.littleblackdogsocialmedia.com/

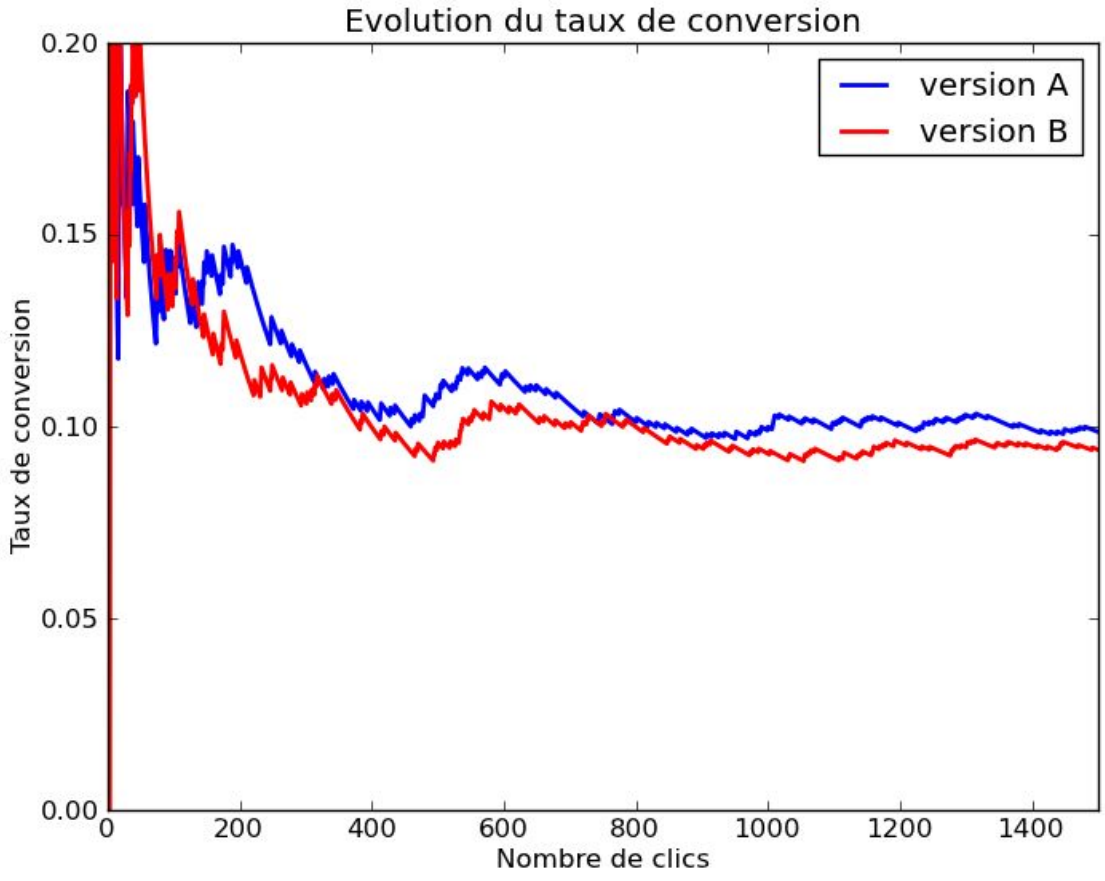
Source : https://vwo.com

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

LANCEMENT DU TEST

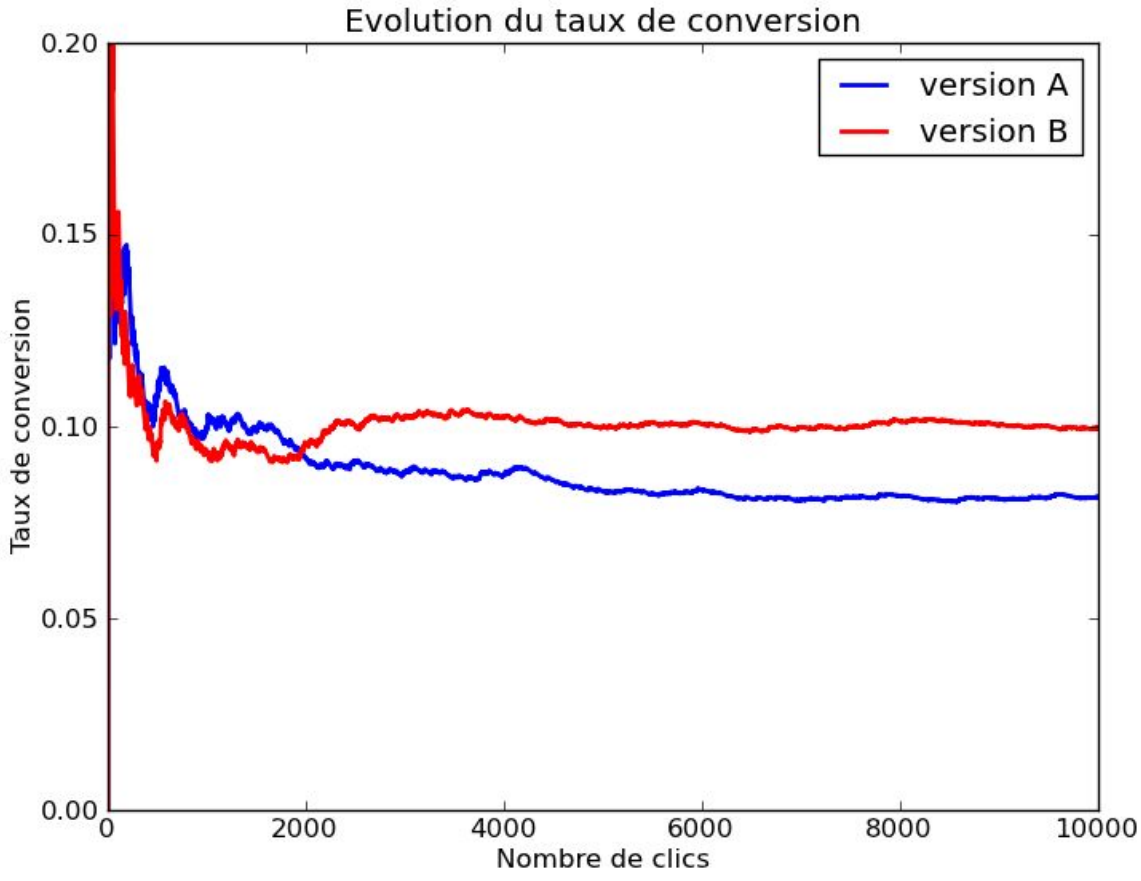


APRÈS 15,000 VISITES



J'implémente la version A !

EN RÉALITÉ



C’était une grosse erreur...

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** vrai ?

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** vrai ?

On ne peut jamais être sûr (à 100%) du résultat. On peut, en revanche, **quantifier la fiabilité** du résultat.

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

ESTIMATION DE LA VARIANCE

Si (X_1, ..., X_n) est un échantillon, on estime sa variance σ^2 par la **variance empirique**:

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

Remarque: La variance représente l'écart moyen à la moyenne.

EXERCICE

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

2, 3, 5, 2, 1

EXERCICE

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

2, 3, 5, 2, 1

moyenne :

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

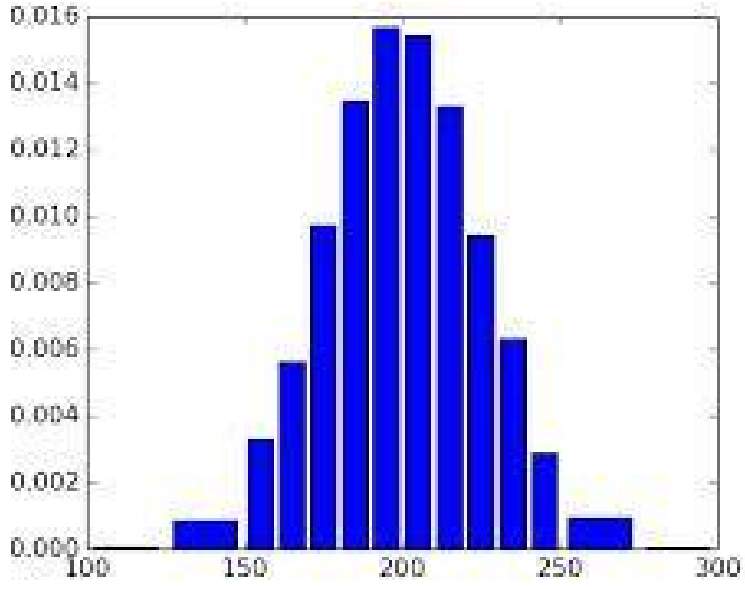
variance :

$$s^2 = 2.3$$

ecart-type :

$$s = \sqrt{2.3} = 1.516$$

RETOUR SUR LA LOI NORMALE



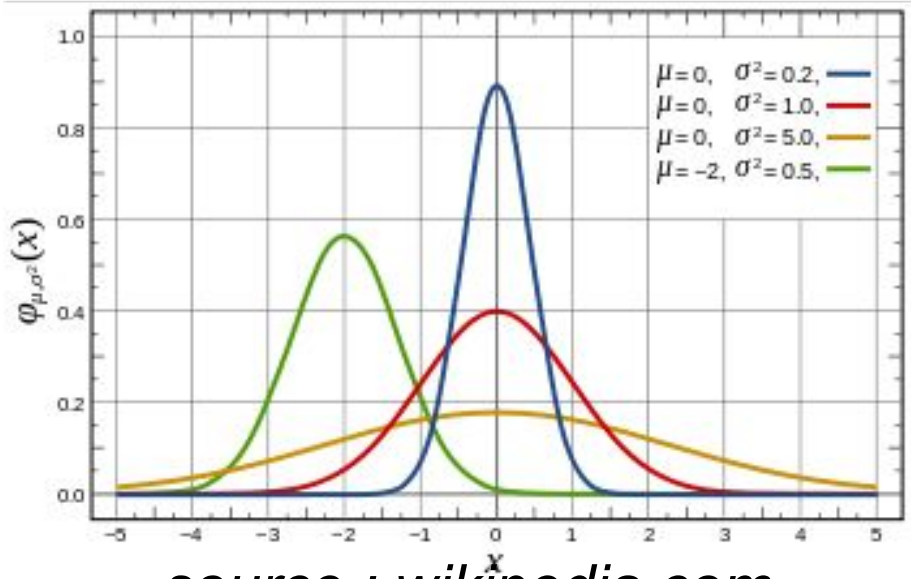
source : matplotlib.org

Histoire de la loi normale :

Loi des erreurs (Gauss, 1777-1855)
L'homme moyen (Quételet, 1796-1874)

L'eugénisme (Galton, 1822-1911)

RETOUR SUR LA LOI NORMALE (2)



source : wikipedia.com

Si X suit la loi normale N (μ, σ_2)

X peut prendre toutes les valeurs entre −1 et +1.

La courbe est symétrique.

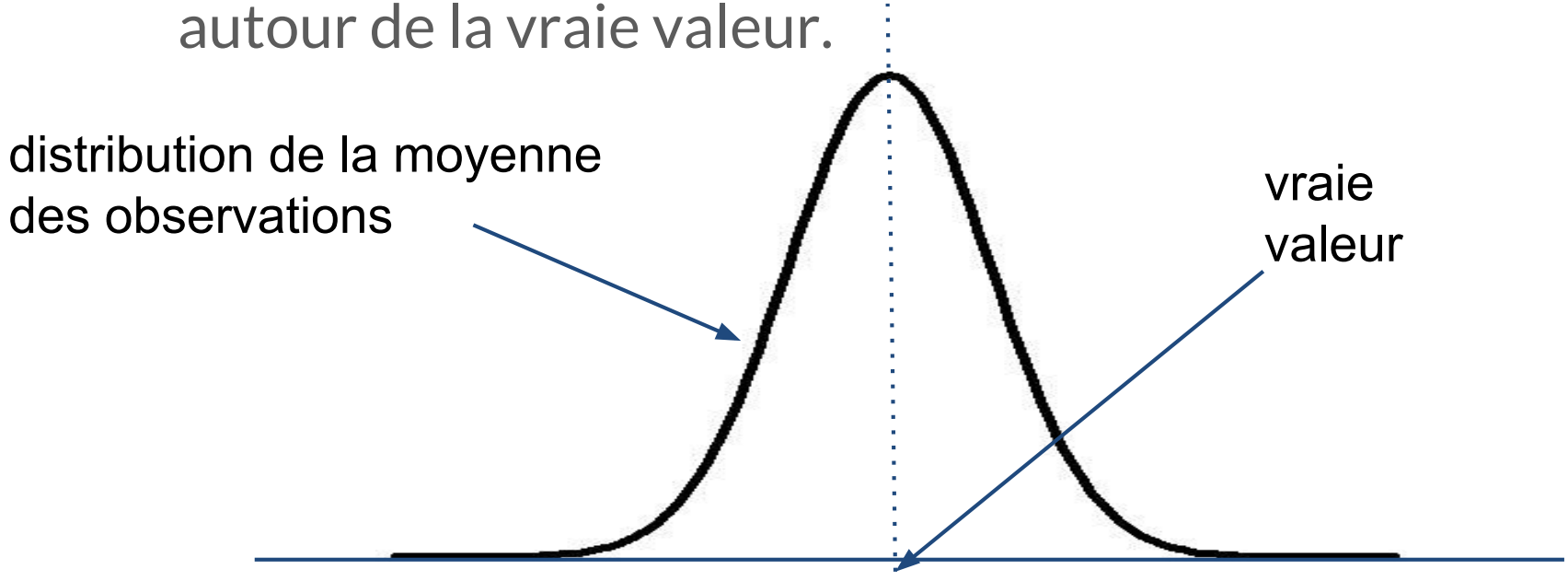
L'espérance de X vaut μ : la courbe est centrée en μ.

La **variance** de X vaut σ_2 (son écart-type vaut donc σ).

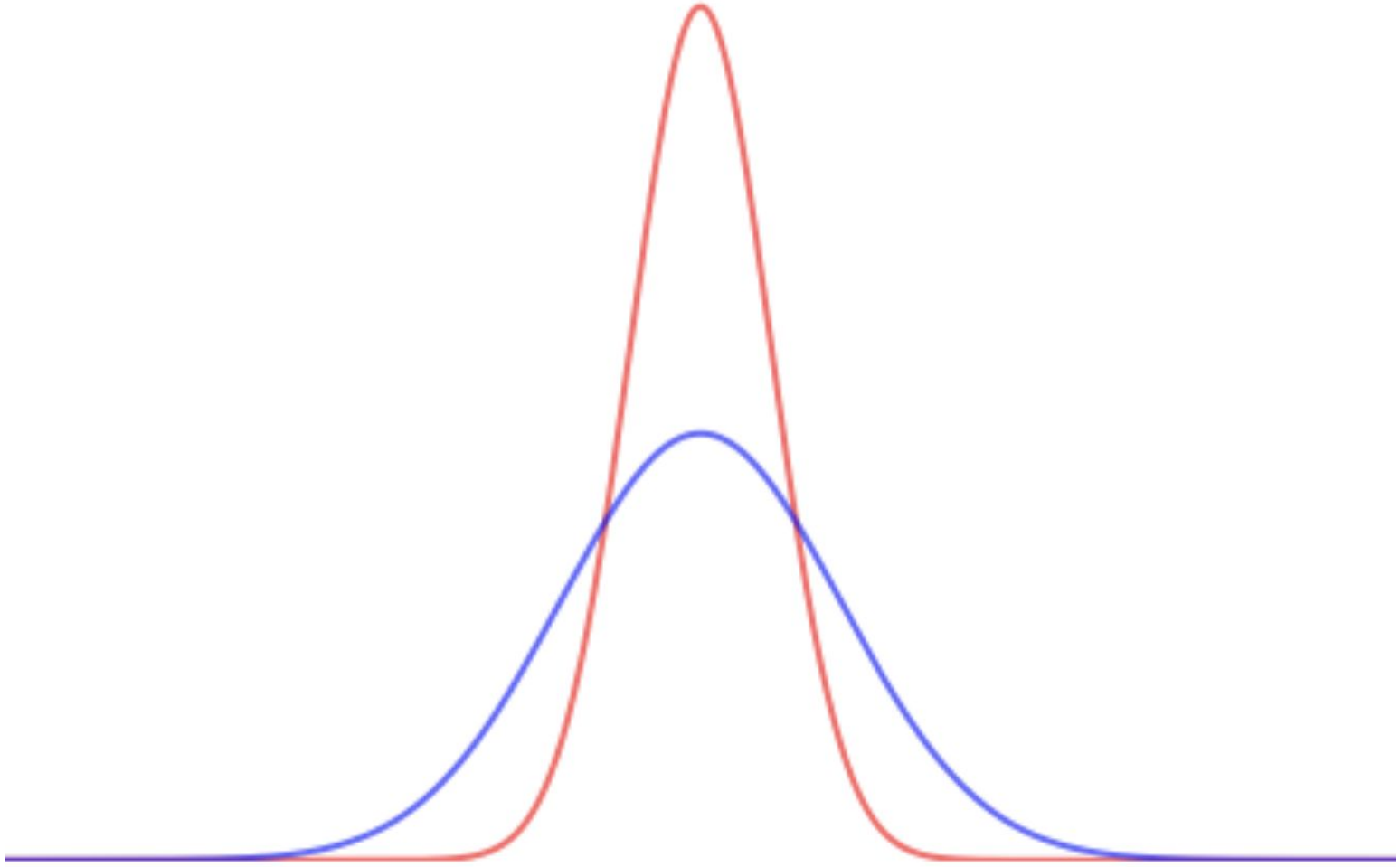
THÉORÈME DE LA LIMITE CENTRALE (1)

Loi des grands nombres : si on fait la moyenne de plusieurs mesures d'une valeur, alors plus on a de mesures, plus cette moyenne s'approche de la vraie valeur.

Théorème de la limite centrale : on sait en plus que cette moyenne des mesures est distribuée selon une loi normale autour de la vraie valeur.

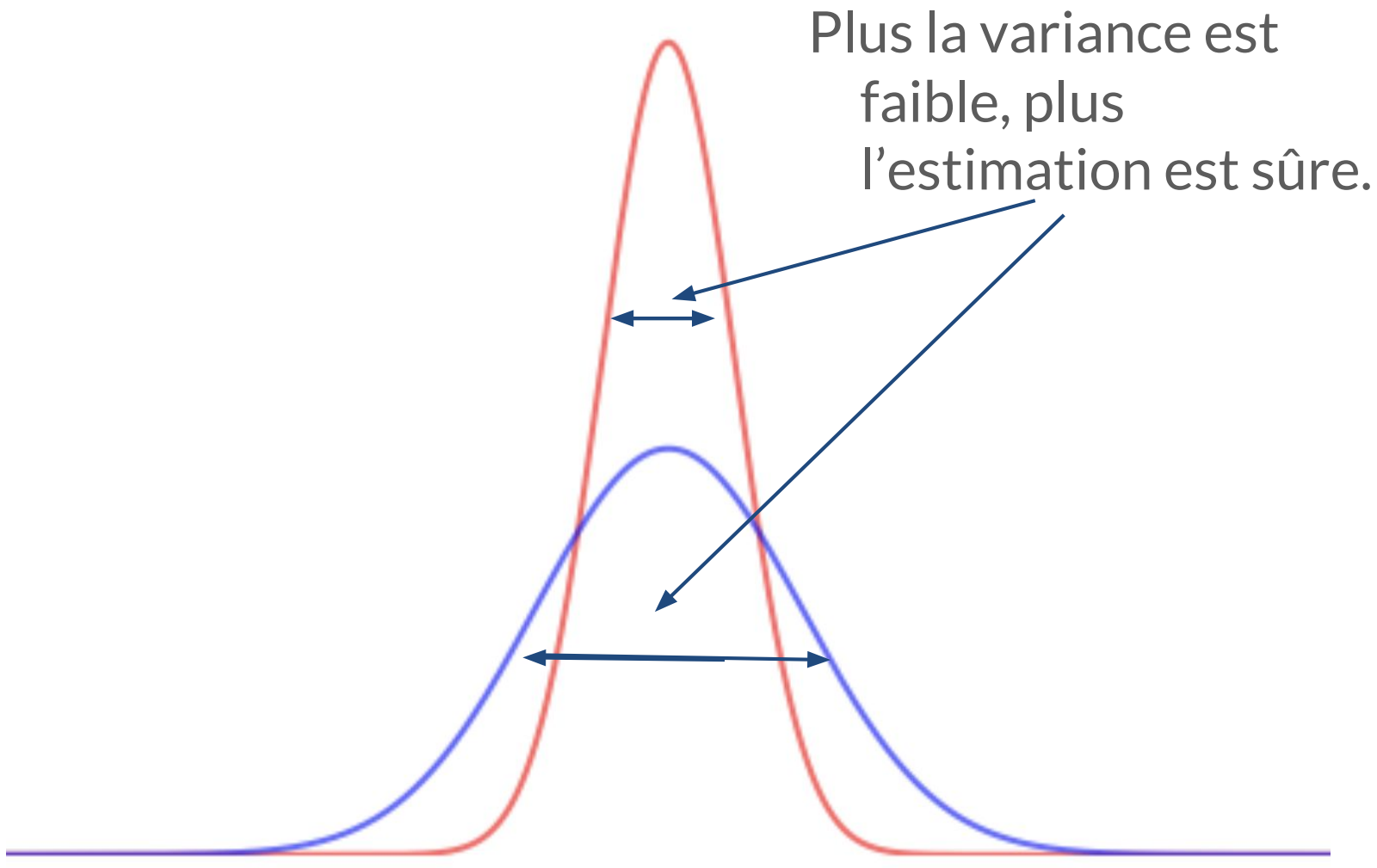


THÉORÈME DE LA LIMITE CENTRALE (2)



Laquelle préfère-t-on ?

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 indépendantes et identiquement distribuées s'approche suit une loi normale de paramètres:

- moyenne : moyenne théorique
- variance : 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 grand résultats statistiques.

Si (X_1, ..., X_n) est un échantillon iid suivant une loi de moyenne μ et de variance σ_2, alors, si n tend vers l'infini, leur **moyenne empirique tend une loi normale** $\mathcal{N}(\mu, \frac{\sigma^2}{n})$:

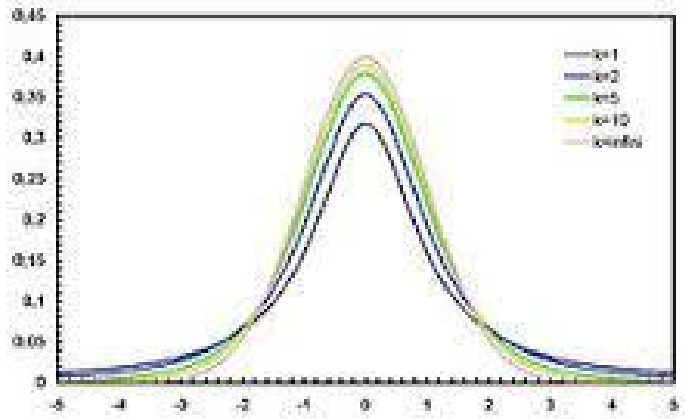
$$\overline{X} \rightarrow_{n \rightarrow +\infty} \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

ou encore: $$\sqrt{n}\frac{\overline{X} - \mu}{\sigma} \rightarrow_{n \rightarrow \infty} \mathcal{N}(0, 1)$$

Remarque: σ^2/n représente la fiabilité de l'estimation. Plus n est grand, plus il est probable que la vraie moyenne μ soit bien approximée par $\overline{X}$

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.



Source : wikipedia.com

Lorsque l'échantillon est suffisamment grand, la loi de Student s'apparente à une loi normale centrée (moyenne 0) réduite (écart-type 1). Pour simplifier, on va admettre cette approximation:

$$\sqrt{n}\frac{\overline{X}-\mu}{s} \rightarrow_{n\rightarrow\infty} \mathcal{N}(0,1)$$

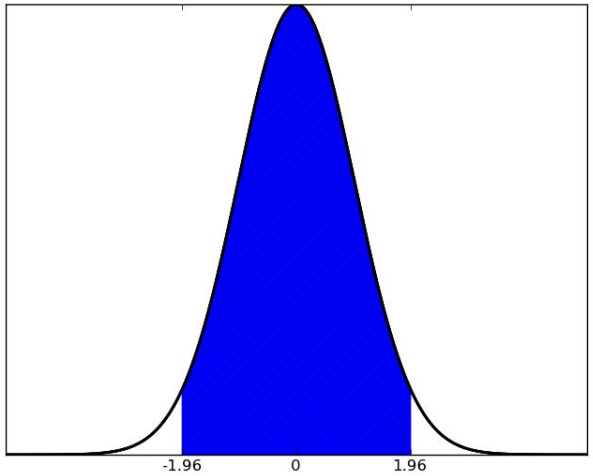
INTERVALLE DE CONFIANCE

Dans le cas d'un échantillon ($X_1 ... X_n$) de moyenne μ, la moyenne théorique se trouve avec une certitude de 95% dans l'intervalle:

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

qui correspond à :

$$\mathrm{P}\left(-1.96 < \sqrt{n}\frac{\overline{x} - \mu}{s} < 1.96\right) = 95\%$$



1.96 : la valeur après laquelle l'aire vaut 2.5%.

−1.96 : la valeur avant laquelle l'aire vaut 2.5%.

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 5% ou plus, il faut dépenser 1 million de plus dans la pub.

Vous observez que sur une semaine, le taux de clics moyen est de 5.1% sur n = 10,000 personnes. Ce qui vous fait un écart-type de 0.22.

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

En remplaçant μ par 0.05, on obtient 0.45, qui est un nombre tout à fait central dans la distribution. Obtenir +/- 0.45 ou une valeur plus extrême a en effet une probabilité de 67% dans une distribution normale centrée-réduite. Vous pouvez prendre votre décision. Inversement, si vous aviez observé une moyenne de 2%, votre statistique vaudrait -21.4. Obtenir +/-21.4 ou une valeur plus extrême a une probabilité qui avoisine 0. Vous ne prenez pas la décision dans ce cas.

LES TESTS STATISTIQUES

Puis-je affirmer que la moyenne μ d'un échantillon vaut la valeur μ_0 avec seulement 5% de chances de me tromper?

Je sais que la moyenne se trouve avec 95% de certitude (donc 5% de risque) entre : $\left[\overline{x} - 1.96\frac{s}{\sqrt{n}}, \overline{x} + 1.96\frac{s}{\sqrt{n}}\right]$

C'est-à-dire que:

$$\mathrm{P}\left(-1.96 < \sqrt{n}\frac{\overline{x} - \mu}{s} < 1.96\right) = 95\%$$

Je calcule $t = \sqrt{n}\frac{\overline{x} - \mu_0}{s}$ car je suppose que μ_0 est la bonne moyenne.

Si t > 1.96 ou t < -1.96 : je suis dans l'erreur, je rejette l'hypothèse et conclue qu'avec 95% de certitude, $\mu \neq \mu_0$.

LES TESTS STATISTIQUES (2)

Si on veut être plus précis:

Si j'affirme que la moyenne μ d'un échantillon vaut la valeur μ_0, avec quelle probabilité ai-je raison?

- Je calcule $t = \sqrt{n}\frac{\overline{X}-\mu_0}{s}$
- Si $t < 0$, je calcule : $2\mathbb{P}(t < \sqrt{n}\frac{\overline{X}-\mu}{s})$
- Si $t > 0$, je calcule $2\mathbb{P}(\sqrt{n}\frac{\overline{X}-\mu}{s} < t)$
- La valeur obtenue s'apelle **p-value**. Plus elle est faible, plus je suis sûre que l'hypothèse est fausse.

LES TESTS STATISTIQUES (3)

Puis-je affirmer que deux moyennes μ_1 et μ_2 sont égales avec seulement 5% de chances de me tromper?

Même principe. Cette fois, il faut calculer:

- $\overline{x}_1, s_1$: moyenne et écart-type empiriques de l'échantillon 1.
- $\overline{x}_2, s_2$: moyenne et écart-type empiriques de l'échantillon 2.
- la variance totale $s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ (la moyenne des deux variances.
- alors, on calcule $t = \frac{\overline{x}_1-\overline{x}_2}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$
- on compare t à -1.96 et 1.96 comme dans l'autre cas.
- si $t > 1.96$ ou $t < -1.96$ on conclue que les deux moyennes sont différentes.
- calcul de la p-value idem que précédemment.

CONCLUSION

Grâce aux **tests statistiques** on peut:

vérifier si une impression est justifiée

donner une conclusion quantifiée en termes de risque

ici on a fait une introduction très superficielle, il existe de nombreux autres tests!