Biostatistiques

I La méthode statistique en médecine :

Les biostatistiques (= statistiques appliquées au domaine de la santé) on 3 objectifs :

1. description d'une population par rapport à une maladie
2. **Evaluation** des traitements, des techniques, des coûts
3. Mise en place des informations épidémiologiques et en tirer des conclusions

Les biostatistiques doivent être capables de décider si une observation peut être due au seul hasard ou si elle a une autre explication.

Quelques définitions de base :

Terme	Définition	Exemple
Statistique	art de collecter, d'analyser, et d'interpréter des données. Lorsqu'elle est appliquée au domaine de la biologie, on parle de biostatistique. Il en existe 2 types : – descriptive – déductive : une observation est-elle due au hasard ? Ou existe-t-il une explication ?	- Stat descriptive : Collecte de 2 données sur la population française : taille et couleur des yeux - Stat déductive : On constate que les sujets ayant une taille > 1,70m ont tous les yeux bleus. Hasard ? Corrélation ?
Population	Série exhaustive de tous les individus étudiés sur lesquels on veut appliquer des décisions	Ensemble de la population française
Echantillon	Ensemble fini et d'effectif limité extrait de la population, le plus souvent randomisé créé par tirage au sort = TAS) donc représentatif	10 personnes tirées au sort dans la population française.
Variable quantitative	résultat de l'observation d'un individu par l'utilisation d'un appareil de mesure, variable d'un individu à l'autre.	Les tailles des 10 individus : 1,62m / 1,63m / 1,66m / 1,66m / 1,68m /1,68m / 1,70m / 1,75m / 1,80m / 1,90m
Variable qualitative	résultat de l'observation d'un individu, par les sens de l'observateur. Variable d'un individu à l'autre	La couleur des yeux : bleus / verts / marrons / gris
Paramètre	grandeur apportant une information résumée sur la variable étudiée	La moyenne : m = 1,708m La médiane : 1,68m
Série statistique	Collection d'objets de même nature, présentant des caractéristiques différentes d'un objet à l'autre	Les hommes et les femmes sont des objets de même nature mais avec des caractéristiques différentes (mince, gros, blond, brun ...)
Variabilité	Ensemble des différences inter-individuelles et intra-individuelles. Elles peuvent être : – dues au hasard – physiologiques	- inter-individuelles : **Comparaison** de la taille des sujets **entre eux.** - intra-individuelles : évolution de la taille avec l'âge, **comparaison du sujet à lui-même** à diverses périodes

II Statistique descriptive :

A Notion de variabilité :

Toute donnée biologique possède une **variabilité**. Sa connaissance est indispensable pour pouvoir dire si la valeur d'une variable **quantitative** est normale ou pas.

- Une variabilité maîtrisée permet d'établir une estimation.
- Une variabilité non maîtrisée conduit à des biais

Exemple :
La **valeur moyenne de la glycémie (variable quantitative)** chez un sujet normal est de **1g/L +/- 0,25 g/L**
Ceci signifie qu'une glycémie appartenant à l'intervalle : **[0,75 g/L ; 1,25 g/L]** est normale.
.
Vos tuteurs de biostat préférés décident de se contrôler la glycémie un matin à jeun. Ils trouvent :

- Vincent : 1,2 g/L. Il a donc une glycémie **normale**
- Julia : 0,7 g/L. Elle a donc une glycémie **infra – normale**

B La représentation des données :

Il existe divers types de données / variables :

- **Les variables qualitatives** :
  - ✗ binaires : homme / femme ; malade / non malade …
  - ✗ nominales : couleur des yeux, des cheveux …
  - ✗ ordinales : degré de satisfaction des étudiants vis à vis de la tut rentrée :
    peu satisfait / satisfait / très satisfait / il n'y a pas de mots tellement c'était bon !

- **Les variables quantitatives** :
  - ✗ discrètes : âge des étudiants en PAES ...
  - ✗ continues : poids, glycémie ...

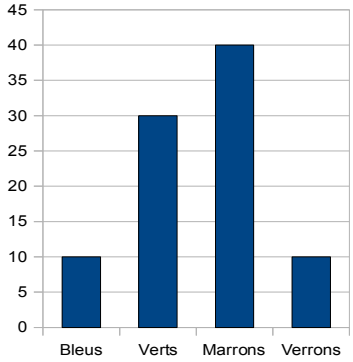
Les variables qualitatives peuvent être représentées de 2 manières :

- diagramme en bâton ou histogramme
- tableau

Exemple : On relève la couleur des yeux de 90 bébés à la sortie de la maternité. On constate que :

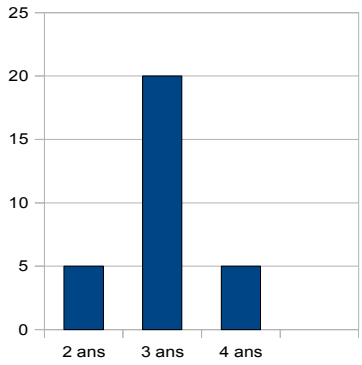
- 10 ont les yeux bleus
- 30 ont les yeux verts
- 40 ont les yeux marrons
- 10 ont les yeux verrons (un oeil vert et l'autre marron)

Couleur des yeux	Nombre de bébés	proportion
bleus	10	11,11%
verts	30	33,33%
marrons	40	44,44%
verrons	10	11,11%



Les variables quantitatives peuvent être représentées/synthétisées de 2 manières :
- diagramme en bâton ou histogramme
- tableau
- synthétisées grâce à des paramètres

Exemple : On relève l'âge de 30 enfants d'une classe de 1ère année de maternelle :
- 5 ont 2 ans
- 20 ont 3 ans
- 5 ont 4 ans



Âge	Nombre d'enfants	proportion
2 ans	5	16,67%
3 ans	20	66,67%
4 ans	5	16,67%

Notion de paramètre : Un paramètre est une grandeur apportant une information résumée sur la variable étudiée. Il en existe plusieurs types :

1) La moyenne :
   - Pour une variable quantitative **discrète** : $\mathbf{m = \sum x_i / n}$
   - Pour une variable quantitative **continue** : $\mathbf{m = \sum n_i\, x_i / n}$

2) La variance : ou (écart type)^2. Elle indique la **dispersion** des données autour de la moyenne.

3) La médiane = **observation centrale des valeurs** qui sépare la série d'un effectif n en **2 sous-séries** de même effectif :
   - Si n est pair, la médiane est donnée par : **n/2**
   - Si n est impair, la médiane est donnée par **(n+1) /2**

4) Les quartiles : Ce sont les valeurs **de la variable** qui partagent la série d'effectif n en **4 sous-séries** de même effectif.

Exemple : On relève le poids de 5 enfants d'une classe d'un service de pédiatrie. On trouve :
- 30 kg
- 31 kg
- 31 kg
- 32 kg
- 33 kg

1) La moyenne : Le poids est une variable quantitative continue.
   m = [(30x1) + (31x2) + (32x1) + (33x1)] / 5 = 31,4kg

2) La médiane : Soient les valeurs suivantes : 30 / 31 / **31** / 32 / 33.
   (n+1) / 2 = 6/2 = 3
   La médiane est donc la **3ème valeur des poids rangés par ordre croissant** : 31 kg.

3) Le 1er quartile : Soient les valeurs suivantes : 30 / 31 / 31 / 32 / 33.
   Le 1er quartile ou Q1 est la **valeur des poids qui délimite les premiers 25%** de la série.
   Q1 = 0,25 x 5 = 1,25 → Q1 se trouve entre la 1ère et la 2ème valeur des poids.
   Soit : (30 + 31) / 2 = **30,5 kg**.

4) Le 3ème quartile : Soient les valeurs suivantes : 30 / 31 / 31 / 32 / 33.
   Le 3ème quartile ou Q3 est la **valeur des poids qui délimite les premiers 75%** de la série.
   Q3 = 0,75 x 5 = 3,75 → Q3 se trouve entre la 3ème et la 4ème valeur des poids.
   Soit : (31 + 32) / 2 = **31,5 kg**.

	avantages	inconvénients
moyenne	– Facile à calculer, à manipuler – significative si la répartition symétrique des données – dispersion faible	– Sensibles aux valeurs anormales – sensible aux minimum – sensible aux maximum
médiane	– Facile à calculer – peu sensible aux valeurs anormales – utilisable pour les valeurs ordinales	– Moins adéquat pour les calculs statistiques

C La notion d'estimation statistique :

1) Généralités :

En biostatistiques, les études sont réalisées sur un **échantillon représentatif** de la population. A l'issue de ces études, se pose le problème de la légitimité des résultats et de leur possible extrapolation à l'ensemble de la population. Pour cela, on réalise une **estimation du résultat vrai**, à partir des résultats obtenus sur l'échantillon.

Il existe 2 types d'estimations :

1. L'estimation ponctuelle : c'est la valeur **unique** jugée la meilleure à l'instant t pour un échantillon donné unique.

2. L'estimation par intervalle : Il s'agit d'un **intervalle de valeurs contenant la valeur recherchée**. Cet intervalle est nommé **intervalle de confiance (= IC)**. L'estimation par intervalle est beaucoup plus fiable que l'estimation ponctuelle.

Afin de consituer un échantillon représentatif de la population, on utilise des techniques précises :

- La détermination précise des caractéristiques de la population
- Le **tirage au sort (TAS)** d'un grand nombre adapté d'individus. +++

L'estimation assure donc la correspondance entre ce qui se passe au niveau de l'échantillon et au niveau de la population.

NB : Deux estimations ponctuelles (respectivement. par intervalle) d'une même variable réalisées sur 2 échantillons A et B donneront des valeurs ponctuelles voisines (resp des IC se recouvrant), mais pas nécessairement la même valeur (resp le même IC)

Exemple : On réalise une étude visant à déterminer la valeur moyenne de la glycémie dans la population française. Pour ce faire, on dispose de 2 échantillons représentatifs A et B d'effectifs $n_A = n_B = 30$ personnes, constitués par TAS. On obtient les résultats suivants :

Echantillon	Echantillon A	Echantillon B
Estimation ponctuelle	Glycémie de 0,95 g/L	Glycémie de 1,03 g/L
Estimation par intervalle	[0,90 – 1,04] g/L	[0,95 – 1,1] g/L

Les estimations **ponctuelles** obtenues sont **voisines** mais non strictement identiques.
Les estimations par **intervalle se recouvren**t.

0,95 g/L (resp. 1,03 g/L) est l'estimation ponctuelle de la valeur moyenne de la glycémie calculée sur l'échantillon A (resp. B).

[0,90 – 1,04] g/L (resp. [0,95 – 1,1] g/L) est **l'IC** calculé sur l'échantillon A (resp. B). La valeur moyenne VRAIE de la glycémie à l'échelle de la population française a donc de forte chances d'appartenir à cet IC.

2) L' estimation de données quantitatives :

L'estimation de données quantitatives se déroule en plusieurs étapes :

1. constitution d'un échantillon représentatif par TAS
2. calcul des paramètres moyenne m et écart type s sur le dit échantillon
3. Estimation de la valeur vraie de la μ moyenne et et de l'écart type σ dans la population

Paramètre	Echantillon	Population
Moyenne	m = estimateur de la moyenne vraie μ au niveau de l'échantillon	μ = moyenne vraie
Ecart type	s = estimateur de l'écart type vrai σ au niveau de l'échantillon	σ = écart type vrai
Effectif	n	N

La notion d'écart type σ:

L'écart type (= √variance) mesure la dispersion d'un ensemble de données autour de la moyenne. Il s'agit donc de la variablité des mesures entre elles et par rapport à la moyenne. Plus il est **faible**, plus le caractère étudié est **homogène** et *vice versa*. Il est donné par la formule :

$$S = \sqrt{(\sum(xi-m)^2/(n-1))}$$

La notion de degrés de liberté :

Les degrés de liberté représentent le nombre des écarts (xi-m) indépendants . Il existe n (x_i-m) et leur somme vaut 0. Il suffit alors de connaître (n-1) x_i pour les connaître tous. Il y a donc n-1 (x_i) indépendants et donc **n-1 degrés de liberté.**

Exemple :

On s'intéresse aux notes obtenues à un tutorat de biostatistiques par 5 élèves de PAES. On récolte les données suivantes :

1. moyenne : 12/20
2. écart type : 13
3. notes obtenues par les 5 élèves : 8, 10, 12, 16, x . La dernière copie ayant été perdue.

notes	8	10	12	16	x
xi-m	-4	-2	0	4	x - 12

Il existe n = 5 notes → il y a 5 (x_i - m). Il y a donc (n-1) = 4 degrés de liberté.
Ces 4 degrés de libertés sont suffisants pour connaître toutes les notes et donc retrouver la note de la copie perdue :

$\sum (x_i - m) = 0$
$- 4 - 2 + 0 + 4 + (x - 12) = 0$
$x - 14 = 0$
$x = 14$

La notion d'intervalle de confiance (=IC) :

L'IC est l'estimation du paramètre m calculé au niveau de l'échantillon mais inconnu au niveau de la population (cf C notion d'estimation). Soit m calculée sur un échantillon. A partir de m, on peut calculer une estimation de la moyenne vraie μ :

μ Є IC
μ Є [m +/- ε s/√n] avec i = εs/ √n = précision de l'IC

- L'IC est aussi appelé intervalle au risque α avec **α** le **risque d'erreur** dans l'estimation de μ, *c'est à dire risque pour l'IC ne contienne pas la vraie valeur de la moyenne μ*, généralement fixé à **5%**.
- **ε** représente l'écart réduit. La valeur d'ε dépend de la valeur d' α (ε et α varient en sens inverse). Pour α = 5% (resp. 1%) , ε vaut 1,96 (resp. 2,6).

Exemple (on gardera le même pour toute l'estimation des données quantatitives):
On réalise une étude visant à déterminer la valeur moyenne de la glycémie dans la population française. Pour ce faire, on dispose d' un échantillon représentatif de n_A = 100 personnes A constitué par TAS. On choisit α = 5%. On obtient les résultats suivants :

α	5,00%	1,00%
Moyenne m sur l'échantillon	m = 0,96 g/L	m = 0,96 g/L
Écart type sur l'échantillon	S = 0,5	S = 0,5
Estimation par intervalle	**IC_{95} = [0,862 – 1,058] g/L**	**IC_{99} = [0,83 – 1,09] g/L**

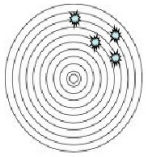
L'IC_{95} est une estimation de μ = la valeur moyenne vraie de la glycémie. Cela signifie qu'il y a **95%** de chances que μ appartienne à l'IC = [0,862 – 1,058] g/L.

L'IC_{99} est une estimation de μ = la valeur moyenne vraie de la glycémie. Cela signifie qu'il y a **99%** de chances que μ appartienne à l'IC = [0,83 – 1,09] g/L.

L'influence d' α sur l'estimation par IC :

Les variations de α conditionnent la précision de l'estimation et la largeur de l'IC :

Large, plus de chances de l'atteindre, mauvaise précision.



Plus **α est petit** et plus l'IC est grand c'est à dire plus ε est grand. *On ne ratera pas la bonne valeur de μ mais la précision sera moindre*.

Exemple : Quand α diminue (α = 1%)
ε augmente (ε = 2,6), donc la largeur de l'IC augmente :
IC_1 = [0,83 – 1,09] g/L.
Il y a **99%** de chances que μ appartienne à l'IC, mais la **précision diminue.**

Resserré, meilleure précision

Plus **α est grand** et plus l'écart réduit ε diminue. L'IC se resserre, *la précision augmente mais la bonne valeur de μ pourra être ratée.*

Exemple : Quand α augmente (α = 5%)
ε diminue (ε = 1,96), donc la largeur de l'IC diminue :
IC_5 = [0,862 – 1,058] g/L.
Il n'y a que **95%** de chances que μ appartienne à l'IC, mais **la précision augmente.**

Taille de l'échantillon et précision :

Plus la taille n de l'échantillon augmente, et plus la précision i augmente c'est à dire que l'estimation tend de plus en plus vers la valeur vraie.

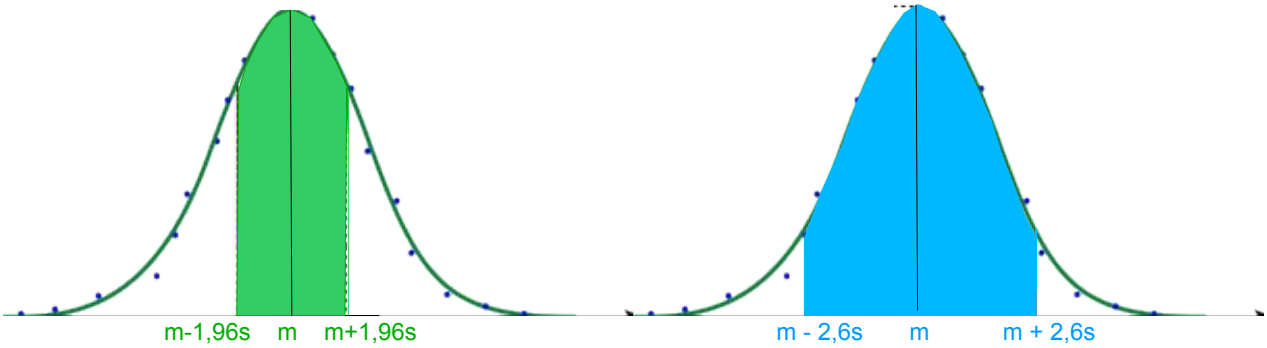
La loi de Gauss ou loi normale :

La loi de Gauss est une loi qui permet, pour tout échantillon où n ≥ 30, de visualiser :

- la notion **d'IC** autour de la moyenne
- la notion **d'écart type**
- la notion **de dispersion** autour de cette valeur moyenne

La **représentation graphique** de données par la loi de Gauss donne une courbe en cloche avec :

1. en abcisse : m +/- εs, donc l'IC
2. en ordonnée : n
3. l'aire sous la courbe : le % de la population concernée



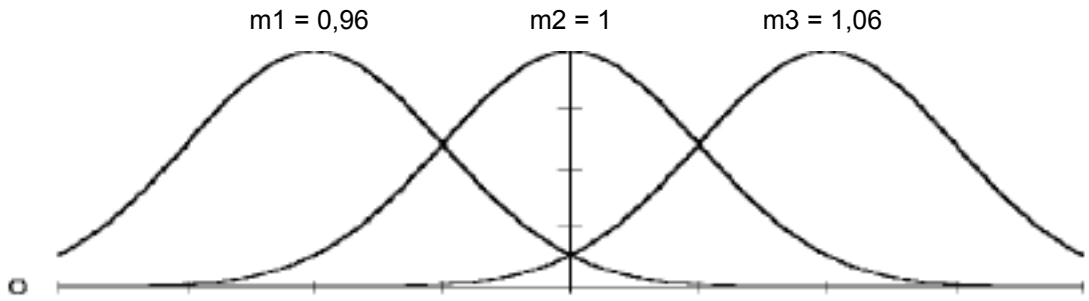
Pour **α = 5%**, ε = 1,96. L'aire comprise sous courbe comprise dans l'intervalle [m – 1,96s ; m + 1,96s] représente **95%** de la population.
Exemple : L'aire comprise sous courbe comprise dans l'intervalle [0,862 ; 1,058] représente 95% de la population.

Pour **α = 1%**, ε = 2,6. L'aire comprise sous courbe comprise dans l'intervalle [m – 2,6s ; m + 2,6s] représente **99%** de la population.
Exemple : L'aire comprise sous courbe comprise dans l'intervalle [0,83 ; 1,09] représente 99% de la population.

Comment la représentation selon la loi de Gauss évolue-t-elle lorsque la moyenne varie ?

Exemple : En réalité, lors de l'étude précédente visant à estimer la valeur moyenne de la glycémie, on avait travaillé, non pas sur un, mais **sur trois échantillons** pour lesquels on obtenait les moyennes suivantes :

Echantillon	Echantillon 1	Echantillon 2	Echantillon 3
m	m1 = 0,96 g/L	m2 = 1 g/L	m3 = 1,06 g/L

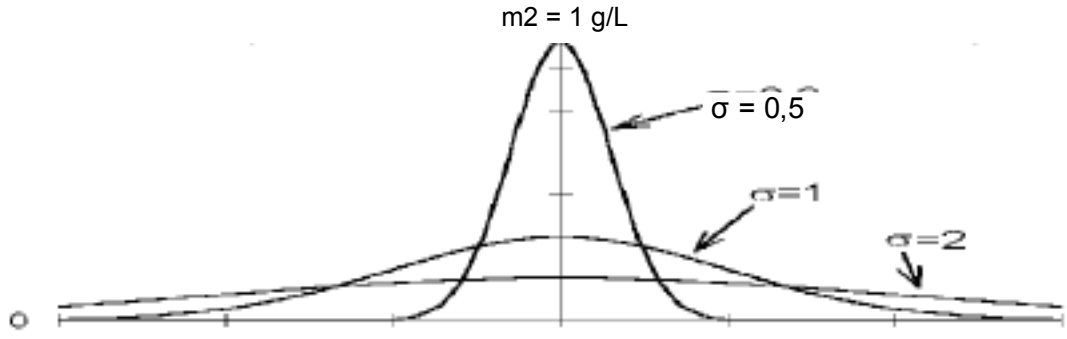


La courbe de gauss est toujours centrée autour de la moyenne m de l'échantillon. Ainsi, on retrouvera exactement la même courbe mais centrée autour de :

- m1 = 0,96 g/L pour l'échantillon 1
- m2 = 1 g/L pour l'échantillon 2
- m3 = 1,06 g/L pour l'échantillon 3

Comment la représentation selon la loi de Gauss évolue-t-elle lorsque l'écart type varie ?

Exemple : On décide, toujours lors de la même étude visant à déterminer la valeur moyenne de la glycémie, de **faire varier l'écart type sur l'échantillon 2**, afin de voir comment il se comporte.



1. Lorsque **σ = 0,5**, la dispersion des données autour de m2 = 1g/L est faible. La proportion des sujets ayant une glycémie proche de m2 tend vers **100%**. Ainsi, on obtient une courbe très "**pointue**".

2. Lorsque **σ augmente (σ = 1 ou 2)**, la dispersion des données autout de la m2 = 1g/L est importante. La proportion des sujets ayant une glycémie très proche de m2 diminue et le proportion des sujets ayant une glycémie éloignée augmentée. Ainsi, la courbe " **s'applatit**"

3) L' estimation de données qualitatives :

Dans l'estimation de données **qualitatives**, on s'intéresse à la proportion de la population présentant une caractéristique quelconque A. Cette estimation se déroule en plusieurs étapes :

1. constitution d'un échantillon représentatif par TAS
2. calcul du pourcentage p_{obs} de l'échantillon présentant A et de l'écart type s
3. Estimation de la valeur vraie p du pourcentage de la population présantant A et de l'écart type σ

Comme précédemment, **l'estimation** assure la correspondance entre ce qui se passe au niveau de l'échantillon et au niveau de la population.

	Echantillon	Population
Proportion (%)	p_{obs} = $p_{observé}$ = estimateur du pourcentage inconnu p	p = pourcentage vrai
Ecart type	s = estimateur de l'écart type vrai σ au niveau de l'échantillon	σ = écart type vrai
Effectif	n	N

La notion d'écart type :
L'écart type a les mêmes caractéristiques pour une variable qualitative que pour une variable quantitative. Il est donné par

$$\mathbf{s = \sqrt{(p_{obs}q_{obs}/n)}}$$

avec $q_{obs} = 1 - p_{obs}$

La notion d'intervalle de confiance :
L'intervalle de confiance est cette fois-ci donné par l'intervalle :

$$\mathbf{p \in IC}$$
$$\mathbf{p \in [\ pobs +/- \varepsilon\ s]\ avec\ i= \varepsilon\ s}$$

L'influence d' α sur l'estimation par IC : idem que pour les variables quantitatives

Précision de l'IC: idem que pour les variables quantitatives
La précision dépend de la taille de l'échantillon n. Plus **l'effectif** de l'échantillon est **grand** et plus la **précision** sera **bonne**.

La précision dépend aussi de l'écart type σ. Quand **s diminue**:

- l'IC diminue
- **la précision augmente**.

Ainsi, si on multiplie n par 100, on multipliera la précision par 10.
A contrario, un IC trop large donnera une mauvaise précision.

Le sondage :
Le sondage est une application directe de l'IC calculée sur des données qualitatives. Les instituts de sondage fournissent toujours sous forme de pourcentage **la valeur centrale de l'IC calculé**. Un IC devrait accompagner tout résultat de sondage.

Exemple :
900 personnes ont été interogées sur leur intention de vote à une élection présidentielle opposant 2 candidats A et B. **52%** ont déclaré qu'ils voteraient pour **A** et **48%** qu'ils voteraient pour **B**. L'institut de sondage annonce que le candidat A arrive en tête avec 52% des voix. Si on accompagne cette valeur centrale de son **IC à 95%** on obtient :
IC = [0,487 ; 0,553]

A récolterait donc, selon le sondage, entre 48,7 et 55,3% des voix. On ne peut donc pas affirmer qu'il arrive en tête.