II-LA STATISTIQUE DESCRIPTIVE

1) Représentation des données

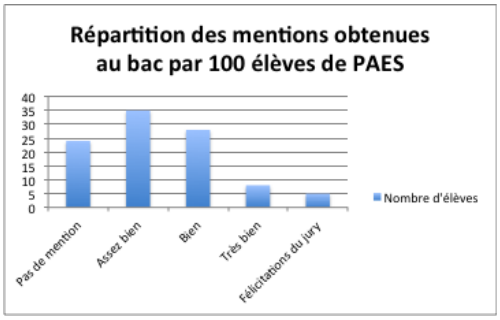
a) Variables qualitatives

Elles peuvent être représentées de deux manières :

- **tableau**
- **histogramme** (normalisé ou non)

Exemple : on étudie les mentions obtenues au bac pour 100 élèves de PAES :

Mentions	Nombre d'élèves
Pas de mention	24
Assez bien	35
Bien	28
Très bien	8
Félicitations du jury	5



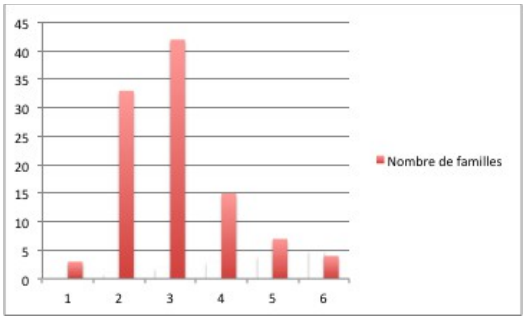
b) Variables quantitatives

Elles peuvent être représentées également de deux manières :

- **tableau**
- **histogramme** (normalisé ou non)

Exemple : on étudie le nombre d'enfants au sein de 100 familles :

Nombre d'enfants par famille	Nombre de familles
0	3
1	33
2	42
3	15
4	7
5	4



Mais pas seulement ! En effet, les variables quantitatives peuvent également être synthétisées ou résumées par des **paramètres** :

=> Soit n données relevées par ordre croissant : x_1 ; x_2 ; x_3 ; x_i ; … x_n

-**la moyenne m** => indicateur de **position**

$$m = \Sigma x_i / n$$

-**la médiane M** : valeur centrale => indicateur de **position**

Effectif n pair	*Effectif n impair*
$M = (x_{n/2} + x_{(n/2)+1}) / 2$ *=>on fait la moyenne de $x_{n/2}$ et $x_{n/2+1}$*	$M = x_{(n+1)/2}$

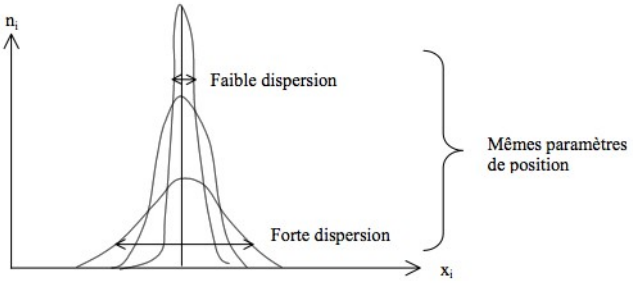
-**les quartiles** : valeurs qui partagent la série ordonnée en 4 groupes de même effectif => indicateur de **position**

*Q1 (premier quartile) sépare les premiers 25% de la série
*Q2 (deuxième quartile) = médiane sépare les premiers 50% de la série
*Q3 (troisième quartile) sépare les premiers 75% de la série

Effectif n multiple de 4	Effectif n non multiple de 4
• **$Q1 = x_{n/4}$** • **$Q3 = x_{3n/4}$**	• **$Q1 = (x_i + x_j) / 2$** avec x_i et x_j les deux valeurs les plus proches de n/4 tq : **$i < n/4 < j$** • **$Q3 = (x_i + x_j) / 2$** avec x_i et x_j les deux valeurs les plus proches de 3n/4 tq : **$i < 3n/4 < j$**

-**la variance (ou écart type)^2** => indicateur de **dispersion**

NB : Pour des courbes tracées suivant les mêmes paramètres de position, on peut rencontrer des formes très différentes. C'est la dispersion qui renseigne sur ces différentes formes, indépendamment de la position des courbes.



Exemple : *On relève le poids de 5 enfants d'un service de pédiatrie :*

- *31 kg*
- *30 kg*
- *31 kg*
- *33 kg*
- *32 kg*

=> La ***série ordonnée croissante*** *est donc :* ***30 / 31 / 31 / 32 / 33***

-*La moyenne* : *(30+31+31+32+33)/5 = 31,4 kg*

-*La médiane* : *valeur centrale de la série ordonnée => 31 kg*

-*Le premier quartile* : *valeur moyenne des poids délimitant les premiers 25% : (30 +31)/2 = 30,5 kg*
(car 5/4=1,25 et 1 <1,25 < 2)

-*Le troisième quartile* : *valeur moyenne des poids délimitant les premiers 75 % : (31+32)/2 = 31,5 kg*
(car 3x5/4=3,75 et 3 < 3,75 < 4)

Il est fondamental de bien distinguer la moyenne de la médiane :

	Avantages	**Inconvénients**
Moyenne	• Facile à calculer • Facile à manipuler => **adaptée aux calculs statistiques** • Significative si la répartition des données est symétrique et la dispersion faible	• Sensible aux valeurs anormales (mini ou maxi)
Médiane	• Facile à calculer • Peu sensible aux valeurs anormales • Utilisable pour les valeurs ordinales	• **Peu adaptée aux calculs statistiques**

NB : La synthèse par des paramètres n'est PAS APPLICABLE pour l'étude de variables qualitatives ! *(par exemple : si l'on rencontre un groupe composé de 4 individus blancs et de 4 individus noirs, on ne peut pas dire que les membres du groupe sont « en moyenne marrons »...)*

2) L'estimation statistique

• **Objectif** : déterminer une grandeur définie sur une population à partir d'observations réalisées sur un échantillon représentatif de cette population. Il existe en fait deux types d'estimations :

• ***estimation ponctuelle*** => valeur jugée la meilleure à un instant t. Estimation très peu fiable.

• ***estimation par intervalle*** => un intervalle de valeurs contenant la valeur recherchée (à un risque d'erreur près) ; c'est ce qu'on appelle l'Intervalle de Confiance (=IC). Beaucoup plus fiable que l'estimation ponctuelle.

• **Méthodologie :**

1) Détermination précise de la population à étudier
2) Tirage au sort d'un échantillon représentatif
3) Etude de l'échantillon
4) Extrapolation des résultats à l'ensemble de la population

=> l'estimation assure à ce niveau la correspondance entre ce qu'il se passe au niveau de l'échantillon et au niveau de la population. Il s'agit la plupart du temps de calculer un Intervalle de Confiance.

Exemple : Une étude vise à déterminer la valeur moyenne de la glycémie dans la population française.

1) Population à étudier : la population française
2) Tirage au sort d'un échantillon A d'effectif connu (par exemple 50 membres)
3) On effectue les prises de sang et le relevé des valeurs
4) L'extrapolation donne :
• estimation ponctuelle => 0,95 g/L
• estimation par intervalle à 95% => [0,90 – 1,04] g/L

A ce stade on peut dire que ***la valeur moyenne VRAIE de la glycémie d'un membre de la population française a donc de fortes chances d'appartenir à cet IC*** *!*

Pour plus de sûreté, on décide de réitérer l'opération sur un nouvel échantillon B tiré au sort, et de même effectif que l'échantillon A. L'extrapolation donne cette fois :
• estimation ponctuelle => 1,03 g/L
• estimation par intervalle à 95% => [0,95 – 1,10] g/L

==> On constate que :
-les ***estimations ponctuelles*** *obtenues* ***sont voisines***
-les ***estimations par intervalle*** *obtenues* ***se recouvrent***

Ce qui va dans le sens d'une confirmation des résultats.

Deux règles primordiales :
-Soient A et B deux échantillons représentatifs d'une même population, alors : • Deux **estimations ponctuelles** d'une même variable réalisée sur les échantillons A et B donneront des valeurs ponctuelles voisines, mais pas nécessairement la même valeur. • Deux **estimations par intervalle** d'une même variable réalisée sur les échantillons A et B donneront des IC se recouvrant, mais pas nécessairement le même IC.

a) Estimation de données quantitatives

ECHANTILLON		POPULATION TOTALE
n = effectif m = moyenne s = écart type	**Estimation**	N = effectif μ = moyenne VRAIE σ = écart type VRAI

On peut calculer les valeurs exactes de la moyenne de l'échantillon (m) et de l'écart type de l'échantillon (s).

Par contre, les valeurs de la moyenne VRAIE de la population μ et de l'écart type VRAI de la population (σ) **ne seront jamais connues** ; on ne peut que les **ESTIMER** !

=> Ainsi, l'Intervalle de Confiance (=IC) est l'estimation de la moyenne VRAIE inconnue de la population (μ) à partir de la moyenne connue de l'échantillon (m). L'IC peut aussi être appelé **Intervalle au risque α**, avec α le risque d'erreur dans l'estimation de μ. Ce risque consiste en ce que l'IC ne contienne pas la valeur vraie de μ. Le plus souvent, on fixe α=5% (5 chances sur 100 que μ soit en dehors de l'IC).

Lorsque l'étude a été correctement menée, on pose ainsi que :

$$\mu \in IC_{(1-\alpha)}$$

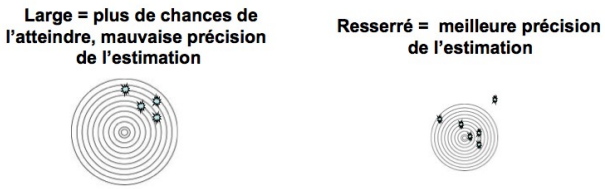
$$IC_{(1-\alpha)} = [\, m - (\varepsilon.s) / \sqrt{n} \;;\; m + (\varepsilon.s) / \sqrt{n} \,]$$

m : moyenne de l'échantillon n : effectif de l'échantillon	s : écart type de l'échantillon	μ : moyenne VRAIE de la population
α : risque d'erreur	ε : écart réduit => *facteur dépendant du risque α*	

Deux valeurs au moins sont à connaître :
-Pour α=5% => ε=1,96
-Pour α=1% => ε=2,6

NB : ε et α varient en sens inverse !

L'IC peut être assimilé à une cible :
-Plus il est grand, plus il contient de valeurs, plus il est susceptible de contenir µ. En contrepartie, sa précision est faible.
-Plus il est petit, moins il contient de valeurs, moins il est susceptible de contenir µ. En contrepartie, sa précision est forte.



Exemple : Une étude vise à déterminer la valeur moyenne de la glycémie dans la population française. Pour cela, on dispose d'un échantillon représentatif de 100 personnes tirées au sort dans la population. On choisit le risque tel que α=5%. Les résultats sont consignés :

α	***5,00%***	***1,00%***
m	*m= 0,96 g/L*	*m=0,96 g/L*
s	*s=0,5*	*s=0,5*
***IC*$_{(1-\alpha)}$**	*IC$_{95\%}$ = [0,862-1,058 g/L]*	*IC$_{99\%}$ = [0,830-1,090 g/L]*

=> Il y a 95% de chance que µ appartienne à l'intervalle [0,862-1,058 g/L].
=> Il y a 99% de chance que µ appartienne à l'intervalle [0,830-1,090 g/L] => moins précis !

C'est l'indice i qui permet de calculer la précision de l'estimation :

$$i = (\varepsilon . s) / \sqrt{n}$$

i : indice de précision s : écart type de l'échantillon
n : effectif de l'échantillon ε : écart réduit => *facteur dépendant du risque α*

NB : la précision AUGMENTE lorsque i DIMINUE.

- Première règle essentielle : LA PRECISION ET LA TAILLE DE L'INTERVALLE DE CONFIANCE VARIENT EN SENS INVERSE !!

• On veut être sur que µ appartienne à l'IC => risque α faible => ε élevé => IC de grande taille => mauvaise précision

• On est prêt à risquer que µ soit en dehors de l'IC => risque α élevé => ε faible => IC de petite taille => bonne précision

- Seconde règle essentielle : LA PRECISION AUGMENTE DANS LE MÊME SENS QUE L'EFFECTIF DE L'ÉCHANTILLON (n).

Une formule permet de calculer le nombre de sujets nécessaires pour une précision et un écart type donnés :

$$n = \varepsilon^2 s^2 / i^2$$

i : indice de précision voulu m : moyenne de l'échantillon s : écart type de l'échantillon
n : effectif nécessaire de l'échantillon ε : écart réduit => *facteur dépendant du risque α*

Etudions désormais une loi exploitant les principaux paramètres des variables quantitatives :

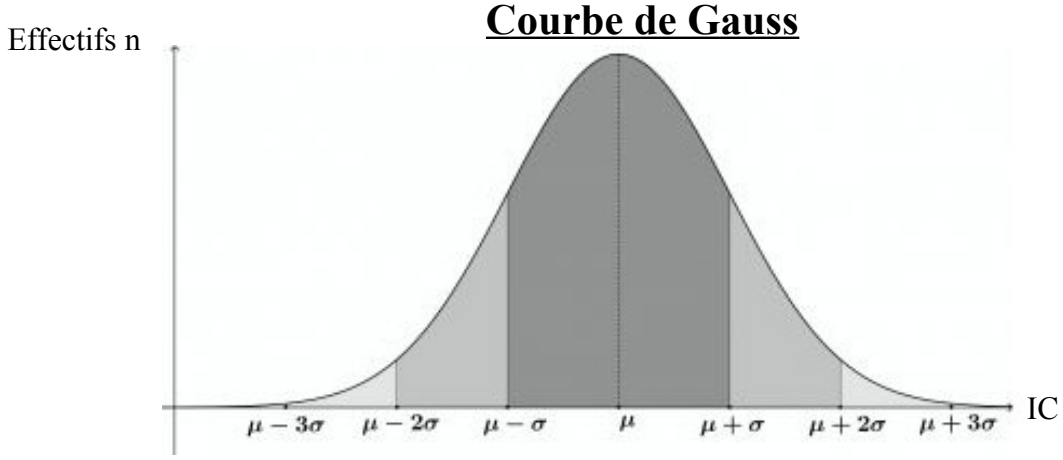
Loi Normale ou Loi de Gauss

Cette loi n'est applicable que pour des populations suffisamment grandes ; c'est à dire avec des effectifs supérieurs ou égaux à 30 !

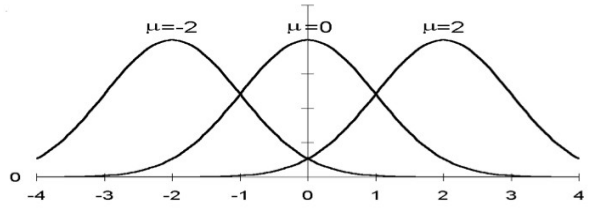
La **Loi Normale** ou **Loi de Gauss** permet de modéliser des phénomènes naturels issus de plusieurs événements aléatoires. On retrouve ainsi sur une courbe en cloche :

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

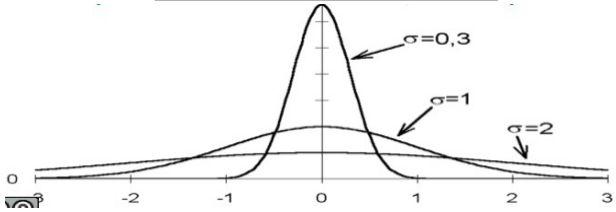
Courbe de Gauss



La valeur de μ modifie la POSITION de la courbe

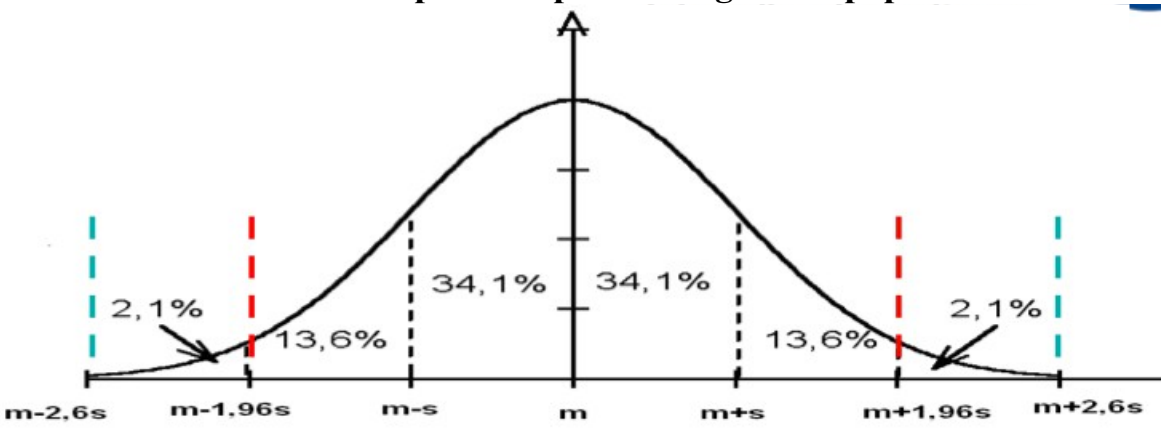


La valeur de σ modifie la FORME de la courbe



La courbe s'aplatit d'autant plus que les valeurs sont dispersée !!!

L'aire sous la courbe correspond au pourcentage de la population concernée



• **IC = [m-1 s ; m+1s]** contient **68,2%** de la population

• **$IC_{95\%}$ = [m-1,96s ; m+1,96s]** contient environ **95,4%** de la population

• **$IC_{99\%}$ = [m-2,6s ; m+2,6s]** contient environ **99,6%** de la population

Exemple : La taille « X » des hommes adultes suit une loi Normale de moyenne ***μ = 180 cm*** *et d'écart type* σ = ***6 cm***.

• La proportion d'homme dont la taille est ***comprise entre*** *174 cm (μ-σ) et 186 cm (μ+σ) est de 34,1% + 34,1% = 68,2%.*

• La proportion d'homme dont la taille est ***inférieure*** *à 168,2 cm (μ - 1,96σ)* ***ou supérieure*** *à 191,8 cm (μ + 1,96σ) est de 2,1% + 2,1% = 4,2%*

(Cette loi sera revue avec d'autres professeurs !)

a) Estimation de données qualitatives

Dans l'estimation de données qualitatives, on s'intéresse à la proportion de la population présentant une caractéristiques quelconque.

ECHANTILLON		POPULATION TOTALE
n = effectif p_o = pourcentage observé s = écart type	**Estimation** →	N = effectif p = pourcentage réel σ = écart type

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

Lorsque l'étude a été correctement menée, on pose ainsi que :

$$\mu \in IC_{(1-\alpha)}$$

$$IC_{(1-\alpha)} = [\ po - (\varepsilon.s)\ ;\ po + (\varepsilon.s)\ ]$$

ou encore

$$IC_{(1-\alpha)} = [\ po - \varepsilon.\sqrt{po.qo/n}\ ;\ po + \varepsilon.\sqrt{po.qo/n}\ ]$$

avec

$$s = \sqrt{po.qo/n} \qquad et \qquad qo = 1-po$$

po : pourcentage observé dans l'échantillon s : écart type de l'échantillon n : effectif de l'échantillon	p : pourcentage réel de la population

α : risque d'erreur ε : écart réduit => *facteur dépendant du risque α*

Deux valeurs au moins sont à connaître :

-Pour α=5% => ε=1,96

-Pour α=1% => ε=2,6

NB : ε et α varient en sens inverse !

• *Exemple 1 : On interroge un échantillon représentatif de 900 personnes au sujet de leur intention de vote à une élection présidentielle opposant le candidat A au candidat B. 52% ont déclaré qu'ils voteraient pour A et 48% se prononcent en faveur du candidat B. Les journaux ont-ils raison d'affirmer que le candidat A arrive en tête du sondage avec 52% des voix ?*

Pour répondre à cette question, on calcule l'IC à 95%, avec ***po=52%*** *et qo=48%*

$$IC_{95\%} = [\ 0{,}52 - 1{,}96\sqrt{0{,}52.0{,}48/1000}\ \ ;\ \ 0{,}52 + 1{,}96\sqrt{0{,}52.0{,}48/1000}\ ]$$

$$IC_{95\%} = [\ 0{,}49\ \ ;\ \ 0{,}55\ ]$$

=> La SEULE conclusion possible est donc que le pourcentage d'intentions de votes pour le candidat A (po) est compris entre 0,49 et 0,55 ; cela avec une certitude de 95%. ***Il est donc FAUX d'affirmer que le candidat A arrive en tête du sondage****, car la valeur de 0,5 est comprise dans l'IC !!!*

• *Exemple 2 : Un célèbre journal affirme que la côte de confiance du président François Hollande a reculé de 2 points à 27% entre juin et juillet. L'article précise que "le sondage a été réalisé du 27 juin au 1er juillet auprès d'un échantillon de 1.000 personnes représentatif de la population âgée de 18 ans et plus".*

Juin 2013		Juillet 2013
• po = 0,29 • qo = 1-po = 0,71 • α = 5% (donc ε = 1,96)	⟶	• po = 0,27 • qo = 1-po = 0,73 • α = 5% (donc ε = 1,96)
$IC_{(95\%)}$ = [0,26 ; 0, 32]		$IC_{(95\%)}$ = [0,24 ; 0, 30]

=> Les IC se recoupent ; ***la chute de la côte de confiance n'est donc, en réalité, pas du tout mise en évidence*** *! Le journal se contentera de dire que celle-ci a chuté de deux points (estimation ponctuelle)…*

Enfin, on peut émettre d'autres réserves au sujet de cet article :

-L'effectif de l'échantillon est-il suffisamment grand ?
-Par quel moyen les participants ont-ils été intérrogés (porte à porte, téléphone, internet…) ?
-La période durant laquelle s'est déroulée le sondage permet-elle une représentativité ?
-Et surtout, y'a-t-il eu des non-réponses ? => La plupart du temps, la somme de po et qo ne vaut pas 1 !!

=> D'où l'importance de TOUJOURS FAIRE ATTENTION AUX CONCLUSIONS D'UNE ENQUETE.

NB : Une NON-REPONSE à un sondage provenant de l'échantillon interrogé constitue toujours un BIAIS !

L'indice i permet de calculer la précision de l'estimation :

$$i = \varepsilon.s = \varepsilon.\sqrt{po.qo/n}$$

i : indice de précision po : pourcentage observé s : écart type de l'échantillon
n : effectif de l'échantillon ε : écart réduit => *facteur dépendant du risque α*

NB : la précision AUGMENTE lorsque i DIMINUE.

-Les **deux règles essentielles** décrites pour l'estimation d'une variable quantitatives s'appliquent ici également.

Calcul du nombre de sujets nécessaires pour une précision donnée :

$$n = \varepsilon^2(po.qo) / i^2$$

i : indice de précision voulu po : pourcentage observé s : écart type de l'échantillon
n : effectif nécessaire de l'échantillon ε : écart réduit => *facteur dépendant du risque α*