# I. LA STATISTIQUE DESCRIPTIVE

Les variables **qualitatives** et **quantitatives** peuvent être représentées de 2 manières :

→ Tableau

→ Histogramme

Mais les variables quantitatives peuvent également être résumées par **des paramètres**

## A. Les Paramètres

Les paramètres ne s'utilisent **que** pour les variables quantitatives !!

Ils permettent de résumer les « caractéristiques » de la série statistique

| **INDICATEURS DE POSITION**<br>*Position des tendances de la série statistique* | **INDICATEURS DE DISPERSION**<br>*Dispersion des données autour d'un indicateur de position* |
|---|---|
| La Moyenne<br>$m = \frac{\sum_{1}^{n} x_i}{n}$<br>*Ex : Sur un groupe de 5 PACES, les notes sont de 12, 8, 10, 5, 14. La moyenne de ces 5 P1 est : (12+8+10+5+14)/5 = 9,8* | La Variance<br>Indique la dispersion des données autour de la moyenne<br>Variance = (Ecart-type) $^2$ |
| La Médiane<br>Elle permet de séparer la série en 2 groupes de **même effectif**<br>C'est la valeur centrale d'une liste ordonnée par ordre croissant<br>→ Effectif n **pair** : $m = \frac{x_{n/2} + x_{(n/2)+1}}{2}$<br>*Ex : 2,5,9,12 → n=4*<br>*n/2 = 2 On fait le moyenne entre la 2ème et la 3ème valeur 5+9/2 =7*<br>→ Effectif n **impair** : $m = \frac{x_{n+1}}{2}$ | L'Ecart-type<br>Moyenne de l'écart à la moyenne<br>Mesure la dispersion des données autour de la moyenne<br>*Ex : Plus les notes des P1 sont homogènes plus la dispersion est faible donc plus l'écart-type est petit* |

### INDICATEURS DE POSITION

Les Quartiles

Les quartiles partagent la série ordonnée en 4 groupes de même effectif

→ **Q1** sépare les premiers **25%** de la série

→ **Q2 (= Médiane)** sépare les premiers **50%** de la série

→ **Q3** sépare les premiers **75%** de la série

- n **multiple de 4** :

☆ Q1 = n/4

☆ Q2 = Médiane

☆ Q3 = 3xn/4

- n **non multiple de 4** :

☆ Q1 = ni+nj/2 *avec i et j : i<n/4<j*

☆ Q3 = ni+nj/2 avec i<3n/4<j

## Moyenne VS Médiane

- La Moyenne :
  - Facile à calculer
  - Adapter aux calculs statistiques
  - Significative si la répartition des données est symétrique et que la dispersion est faible (petit écart-type)
  - Sensible aux valeurs anormales
- La Médiane :
  - Facile à calculer
  - Peu sensible aux valeurs anormales
  - Utilisable pour les valeurs ordinales
  - Peu adaptée aux calculs statistiques

## B. L'estimation statistique

<u>Objectif</u> **: On veut déterminer une grandeur définie sur une population à partir d'observations réalisées sur un échantillon de cette population**

*<u>Ex</u> : durée moyenne d'une pause pour un P1 à la bibliothèque SJA à Nice*

Pour cela il y a deux types <u>d'estimation statistique</u> :

→ Estimation **PONCTUELLE** : Valeur, jugée la meilleure à un instant t (peu fiable)

→ Estimation par **INTERVALLE** : Intervalle de valeurs contenant la valeur recherchée = <u>intervalle de confiance</u>

### 1) La Méthodologie

- On détermine précisément la population à étudier = **population cible**
- Tirage Au Sort de n sujets = échantillon **représentatif**
- Etude <u>sur</u> l'échantillon → **Calcul de l'intervalle de confiance (estimation)**

  → **Extrapolation à la population**

*<u>Exemple</u> : On mesure la glycémie de 2 échantillons représentatifs de la population des P1 de Nice*

| | ECHANTILLON A | ECHANTILLON B |
|---|---|---|
| ESTIMATION PONCTUELLE | 0,89 g/L | 0,97 g/L |
| ESTIMATION PAR INTERVALLE | [0,83 g/L ; 0,98 g/L] | [0,92 g/L ; 1,04 g/L ] |

Deux **estimations ponctuelles** d'une même variable réalisées sur les échantillons A et B donneront des **valeurs ponctuelles voisines**, mais pas nécessairement les mêmes valeurs

Deux **estimations par intervalles** d'une même variable réalisées sur les échantillons A et B donneront des **Intervalles de confiance (IC) qui se recouvrent**, mais pas nécessairement les mêmes.

### 2) L'Intervalle de confiance

La moyenne **vraie** de la **population** et l'écart-type ne peuvent être connus, mais on peut connaître les paramètres sur l'échantillon : on **ESTIME** ceux de la population

- Pour des données **quantitatives**

| ECHANTILLON | Estimation → | POPULATION TOTALE |
|---|---|---|
| n = effectif<br>m = moyenne<br>s = écart type | | N = effectif<br>μ = moyenne VRAIE<br>σ = écart type VRAI |

**α : Probabilité de se tromper dans l'estimation de μ** c'est à dire que l'IC ne contienne pas la valeur vraie de μ !!

**ε : écart réduit**

$$\mu \in \left[ m \pm \frac{\varepsilon s}{\sqrt{n}} \right] \Rightarrow \text{Intervalle au risque } \alpha$$

L'estimation assure la correspondance entre l'échantillon et la population

α= 5% → ε = 1.96 Si α diminue ε augmente !!

α = 1% → ε = 2.6

➢ <u>Indice de précision</u> : Cette valeur est la **largeur** de l'intervalle de confiance

PLUS L'INDICE EST <u>PETIT</u>
<u>MEILLEURE</u> EST LA PRECISION !!

$$i = \varepsilon \frac{s}{\sqrt{n}}$$

➢ <u>Nombre de sujets nécessaires</u> :

$$n = \varepsilon^2 \frac{s^2}{i^2}$$

## PROPRIETES A CONNAITRE ♡♡ +++++++

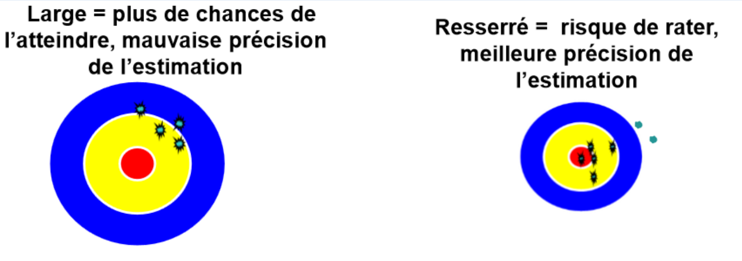
*Regarder bien les formules pour comprendre !!!*

Plus α est petit, plus l'intervalle est grand !!

Si la taille de l'échantillon augmente, la précision augmente !! Donc i diminue

Plus l'IC est large, moins il est précis !!

*L'IC peut être assimilé à une* ***cible*** *:*



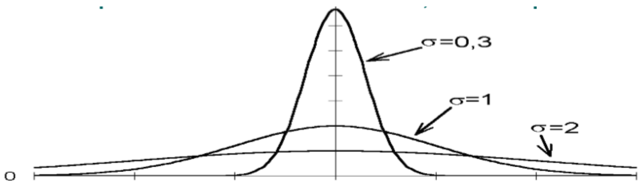

LA PRECISION ET LA TAILLE DE L'INTERVALLE DE CONFIANCE VARIENT EN SENS INVERSE !

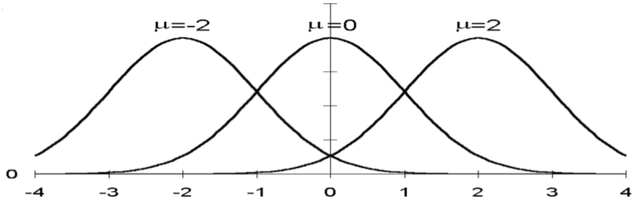
Si l'effectif n augmente → IC se resserre → indice i diminue → précision augmente

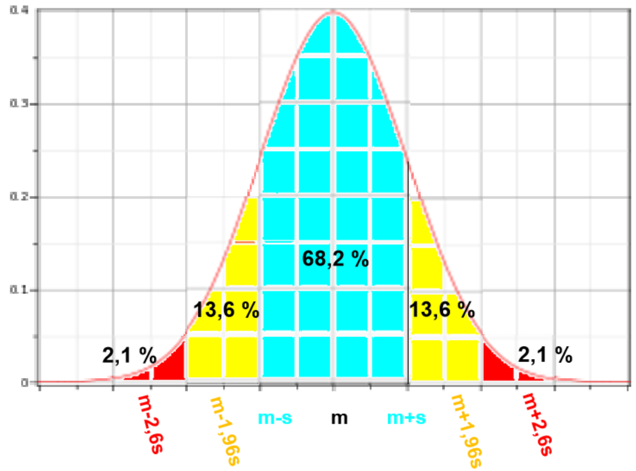
- La **Loi Normale** ou **courbe de Gauss**

Sur cette courbe en cloche on retrouve :

- L'écart-type σ : la dispersion autour de la moyenne μ




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

Cette loi ne s'applique que pour des effectifs de **plus de 30 personnes** !

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

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

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

- Pour des données **qualitatives**

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

$p_{obs}$ : Estimateur du pourcentage inconnu p

Estimateur de l'éart-type inconnu :

$$\boxed{s = \sqrt{\frac{p_0 q_0}{n}} \quad \text{avec } q_0 = 1 - p_0}$$

$$p \in [\, p_{obs} - \varepsilon s \,;\, p_{obs} + \varepsilon s \,]$$

- <u>Indice de précision</u> : $$i = \varepsilon \sqrt{\frac{p(1-p)}{n}}$$

**Rappel : Plus i est petit meilleure est la précision !**

- <u>Nombre de sujets nécessaires</u> (NSN) : $$n = \varepsilon^2 \frac{p(1-p)}{i^2}$$

n multiplié par 100 → s divisé par 10 → précision augmente facteur 10

*<u>Exemple</u> :* Pour décider si oui ou non des télés seront mises en place les chambres d'un hôpital, un sondage est effectué auprès d'un échantillon représentatif de 200 malades. 110 personne souhaitent leur mise en place, les 90 autres ne trouve pas d'intérêt à avoir des télés. L'hôpital peut-il commencer à acheter les tv ?
P0= 110/200 = 0.55
IC95% = [0,55 – 1,96√0,55*0,45/200 ; 0,55 + 1,96√0,55*0,45/200]
IC95% =[0,48 ; 0, 62] L'intervalle de confiance à 95% indique des possibilités de valeurs **en dessous de la moyenne**. L'hôpital ne peut donc rien conclure de ce sondage

→ TOUJOURS FAIRE ATTENTION AUX CONCLUSIONS D'UNE ENQUETE

→ Une NON-REPONSE à un sondage provenant de l'échantillon interroge constitue toujours un BIAIS !



# II. LA STATISTIQUE DEDUCTIVE

*Rappel : tirer des conclusions à partir d'observations*
*<u>Exemple</u> : Comparer 2 groupes pour un caractère donné*

## 1) Les Hypothèses

Il y a 2 hypothèses :

→ **H0 : hypothèse nulle** : il n'y a pas de différence observée entre les deux groupes

→ **H1 : hypothèse alternative** : il y a une différence significative entre les deux groupes

On choisit toujours pour **H0** l'hypothèse qu'il serait <u>**le plus grave de rejeter à tort**</u>
Les 2 hypothèses jouent des rôles symétriques !

Les tests permettent de décider si on accepte ou on rejette H0 au risque $\alpha$

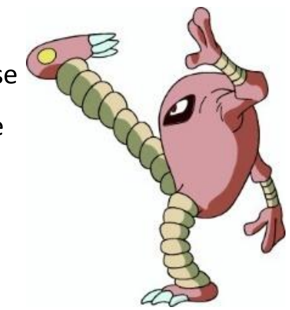
**Décision du statisticien**

| Réalité | Rejet H0 | Non rejet H0 |
|---|---|---|
| H0 Vraie | Erreur 1ère espèce $\alpha$ | $1 - \alpha$ |
| | | |
| H1 Vraie | Puissance $1 - \beta$ | Erreur 2ème espèce $\beta$ |

☆Risque de **1ère espèce $\alpha$** : Probabilité de rejeter H0 si H0 vraie

☆Risque de **2nde espèce $\beta$** : Probabilité d'accepter H0, si H0 fausse

☆**Puissance du test** $1 - \beta$ : Probabilité de rejeter H0 si H0 fausse



## 2) Les Etapes de mise en œuvre d'un test d'hypothèse

- **Etape 1** : Avant recueil des données **définir H0 et H1**
- **Etape 2** : Définir le test en fonction du **type des données** (qualitatives, quantitatives). Soit Z le paramètre qui sera calculé
- **Etape 3** : Choisir le risque $\alpha$ (dans la pratique souvent 5%)
- **Etape 4** : Recueil des données + Calcul de Z
  Règle de décision : examiner la position de cette valeur Z, par rapport à un modèle théorique dont on connait la distribution.
- **Etape 5** : Interprétation des résultats : *Accepte-t-on H0 au niveau de l'échantillon ? Peut-on extrapoler à la population ?*

## 3) Les Tests

### Nombre de degrés de libertés (ou ddl)

*Le nombre de degré de liberté se traduit par le nombre minimal de données qu'il est nécessaire de connaître afin de pouvoir déduire toutes les données manquantes*

- leur somme =0
- il suffit d'en connaitre (n-1) pour les connaitre tous : n-1 degrés de liberté
- Il y a n écarts (Xi – m)

*(On se calme ça fait peur comme ça, mais il faut juste connaître les formules des ddl dans les différents tests)*

- Etude de la liaison entre deux caractères **qualitatifs**

On peut choisir d'utiliser :

→ **Un test de comparaison des pourcentages :**

On réalise les 5 étapes citées ci-dessus :

1) **H0 =** il n'y a pas de différence significative entre les deux groupes *(ex : Pas plus d'yeux bleus dans un groupe que dans l'autre)*
**H1 =** il y a une différence significative entre le groupe A et B c'est à dire que la proportion d'individus du groupe A présentant x est différente de celle du groupe B.

2) Les variables sont :
• des types d'individus : variable **qualitative**
• une caractéristique x **qualitative**
→ Variables qualitatives : on peut donc choisir le test de comparaison des %

3) On choisit le seuil d'erreur de 1 ère espèce α généralement fixe à 5%.

4) Recueillir les données, calculer Z et utiliser la règle de rejet
La variable Z est ici représentée par **l'écart réduit ε**
Ici on compare donc :
- **l'écart réduit ε théorique** donné par la table de l'écart réduit en fonction de $\alpha$
- **l'écart réduit ε calculé** : $\dfrac{p_A - p_B}{\sqrt{\dfrac{p_A q_A}{n_A} + \dfrac{p_B q_B}{n_B}}}$
(Pas à apprendre)

5) Si **ε théorique** > **ε calculé** alors on accepte H0 et on rejette H1
Si **ε théorique** < **ε calculé** alors on accepte H1 et on rejette H0

→ **Un test du X² (chi²)**
*même méthode que ci-dessus*

- Le **X² théorique** est donnée par la table du X² : intersection entre le nombre de ddl et α
- Pour le test du X², le nombre de ddl vaut : ($n_{lignes}$ – 1) x ($nb_{colonnes}$ – 1)
- X² calculé = $\sum \dfrac{(Oi - Ci)^2}{Ci}$ (Pas à apprendre)

Si **X² calculé** > **X² théorique** : on rejette H0 : on accepte H1 au risque α (on conclut qu'il y a une différence)

Si **X² calculé** < **X² théorique** : on accepte H0 au risque α (il n'y a pas de différence)

- Etude de la liaison entre caractères **qualitatifs et quantitatifs**

On peut choisir d'utiliser :

**→ Un test de comparaison des moyennes** :
En présence de données qualitatives et quantitatives, pour **n1 et n2 > 30**, on peut choisir d'utiliser un test de comparaison des moyennes

- La variable Z est ici représentée par **l'écart réduit ε**
- **ε théorique** est donné par la table de l'écart réduit, en fonction de α
- **ε calculé** : $= \dfrac{m_1 - m_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

(Pas à apprendre)

Si **ε calculé** > **ε théorique** on rejette H0 et on accepte H1
Si **ε calculé** < **ε théorique** on rejette H1 et on accepte H0

Si le / les échantillon(s) considérés sont représentatifs de la population, on pourra alors extrapoler le résultat obtenu à l'ensemble de la population.

**→ Un test du t de student** :
En présence de données qualitatives et quantitatives, pour **n1 ou n2 < 30**, on peut choisir d'utiliser un test du t de student

- le **T student théorique** est donné par la table du T student : intersection entre ddl et α
- Nb ddl : (n1 – 1) + (n2 - 1)
- Le **T student calculé** = $\dfrac{m_1 - m_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

Si **T student calculé** > **T student théorique** on rejette H0 et on accepte H1
Si **T student calculé** < **T student théorique** on rejette H1 et on accepte H0

- Etude de la liaison entre deux caractères **quantitatifs**

**→ Le coefficient de corrélation** :

- On recueille différentes valeurs de x et de y.
- On trace la courbe y = f(x)
- **La pente** de cette droite est appelé **coefficient de corrélation r**

Il est donné par la formule suivante :
(Pas à apprendre)

$$r = \frac{\sum xy - \dfrac{\sum x . \sum y}{n}}{\sqrt{\left(\sum x^2 - \dfrac{(\sum x)^2}{n}\right)\left(\sum y^2 - \dfrac{(\sum y)^2}{n}\right)}}$$

- r est toujours compris entre [-1 ;1]
- ddl = n-2

Si il **n'existe pas ou est nul**, alors il n'y a **pas de corrélation** entre x et y au niveau de l'échantillon
si r existe et **r > 0**, alors il existe une **corrélation positive** entre x et y au niveau de l'échantillon
si r existe et **r < 0**, alors il existe une **corrélation négative** entre x et y au niveau de l'échantillon

Au niveau de l'échantillon si **|r calculé| > |r théorique|** trouvé dans la table de r : alors, on **rejette H0**