

UBS

ETUDE DE CAS

**Analyse de l’initiation de la marche chez des enfants âgés de 6 à 12 ans, neurotypiques et atteints de trouble du spectre de l’autisme**

*Effet du TSA*

Auteurs

Daouda diop

Brun Bahoun Houtoukpe

## Introduction

Dans le cadre de cette étude, nous nous sommes intéressés à l’analyse de l’initiation de la marche chez des enfants âgés de 6 à 12 ans, à travers la comparaison entre un groupe d’enfants neurotypiques (NT) et un groupe d’enfants présentant un trouble du spectre de l’autisme (TSA). Le trouble du spectre de l’autisme se caractérise par une grande hétérogénéité des symptômes et une variété d’intensités, notamment en matière de communication, de comportements sociaux et, dans certains cas, de capacités motrices.

L’objectif principal de cette seconde partie était de **comparer les deux groupes** sur plusieurs aspects (données générales, paramètres biomécaniques, scores cliniques), d’**explorer les relations entre variables motrices et scores TSA**, et d’**identifier d’éventuels profils d’enfants similaires** à l’aide d’une approche de clustering non supervisé.

Toutes les analyses ont été réalisées avec le logiciel R. Le code correspondant ainsi que les fichiers de données ont été fournis en annexe. Ce rapport présente, pour chaque question posée, les traitements réalisés, les justifications méthodologiques des choix effectués (tests statistiques, visualisations, regroupements) ainsi que les résultats obtenus sous forme de tableaux et de graphiques.

## Q2.1 Appariement des enfants NT et TSA

### **1. Comment a été effectué l’appariement ?**

Dans cette étude, il a été nécessaire de sélectionner **30 enfants neurotypiques (NT)** parmi **40 initialement disponibles**, afin de les apparier avec **30 enfants atteints de trouble du spectre de l’autisme (TSA)** en fonction de trois critères :

* **L’âge** (Age),
* **La taille** (Height),
* **Le poids** (Weight).

Plusieurs approches peuvent être envisagées pour réaliser un appariement entre deux groupes :

1. **Appariement exact** :  
   Cette méthode consiste à sélectionner les enfants NT ayant exactement les mêmes valeurs d'âge, de taille et de poids que les enfants TSA. Toutefois, cette approche est souvent **trop restrictive**, car elle limite fortement les possibilités d’appariement et risque d’exclure de nombreux sujets.
2. **Appariement par le plus proche voisin ("Nearest Neighbor Matching")** :  
   Cette technique consiste à comparer chaque enfant TSA aux enfants NT disponibles et à sélectionner celui dont les caractéristiques sont les plus proches, en minimisant la **distance** entre les valeurs des trois critères d’appariement.
   * Cette distance est généralement calculée à l’aide de la **distance euclidienne** ou du **score de propension**.
   * Une seule correspondance est attribuée à chaque enfant TSA.
3. **Appariement avec une marge de tolérance ("Caliper Matching")** :  
   Cette méthode impose une **marge de tolérance** (appelée "caliper") pour limiter l'écart entre les valeurs des variables d’appariement, garantissant ainsi une correspondance plus stricte. Cependant, si les données ne sont pas suffisamment nombreuses, cette méthode peut **réduire le nombre d’individus appariés**.

### **2. Justification du choix de l’appariement par le plus proche voisin**

L’appariement a été réalisé en utilisant la méthode du **plus proche voisin ("Nearest Neighbor Matching")** pour plusieurs raisons :

* **Souplesse** : Contrairement à l’appariement exact, cette approche permet d’obtenir un appariement optimal sans exclure trop d’individus.
* **Précision** : Elle minimise l’écart entre les enfants NT et TSA en optimisant la similarité sur les critères définis (âge, taille et poids).
* **Fiabilité statistique** : Elle permet d’obtenir deux groupes équilibrés, réduisant ainsi les biais dans l’analyse des résultats.

Cette méthode garantit un échantillon **comparatif équilibré**, essentiel pour analyser les effets du TSA sans introduire de biais liés aux caractéristiques démographiques et anthropométriques.

### **3. Comment vérifier que l’appariement a été correctement effectué ?**

Pour s'assurer que l’appariement est bien réalisé, plusieurs analyses ont été effectuées :

##### **a) Comparaison des statistiques descriptives entre NT et TSA**

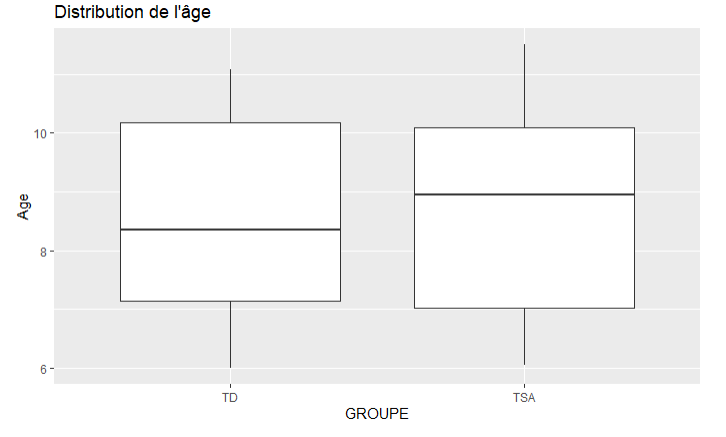
Avec R, nous avons calculé les **moyennes et écarts-types** des variables **âge, taille et poids** pour chaque groupe :



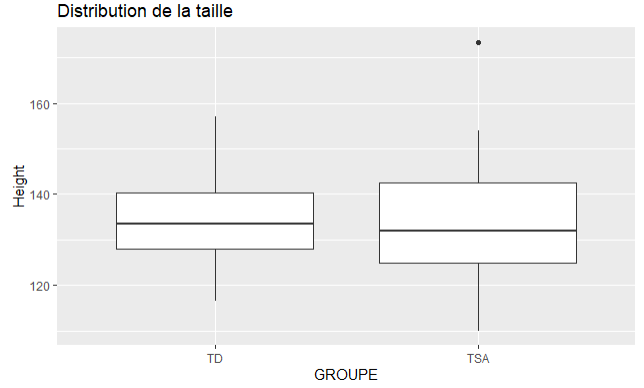
Les moyennes et les écarts-types sont proches entre NT et TSA, indiquant un bon équilibre entre les groupes.

### b) Visualisation des distributions

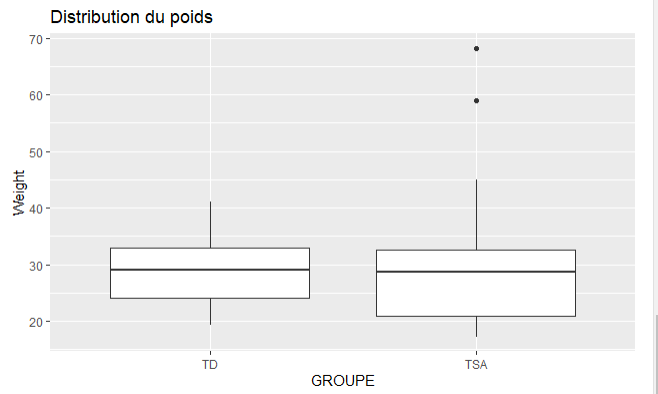
Nous avons également réalisé des boxplots pour visualiser la répartition des variables.



Les distributions sont similaires entre les deux groupes.



Au niveau de la taille, il y a quelques variations mais pas de différences majeures entre les groupes.



Pour le poids, il y a quelques valeurs extrêmes détectées, mais la distribution générale est homogène.

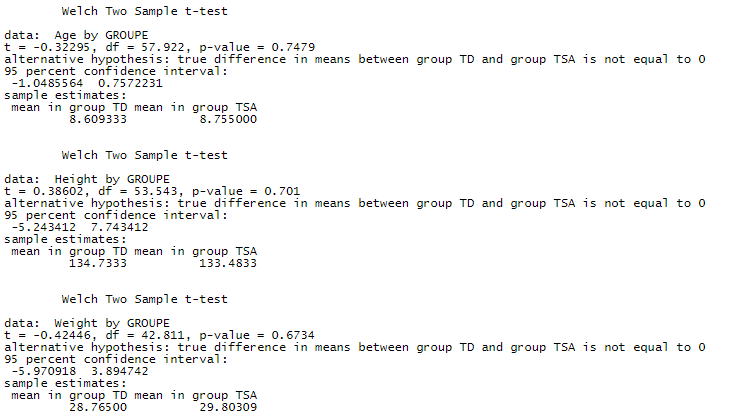
Les boxplots confirment que les distributions des variables sont comparables entre les groupes NT et TSA.

### c) Tests statistiques pour valider l’appariement

Nous avons réalisé deux types tests de comparaison des groupes :

* **Test de Student (t-test)** pour comparer les moyennes
* **Test de Mann-Whitney (Wilcoxon Rank Sum Test)** pour vérifier si les distributions sont similaires

**Résultats des test de Student avec R**

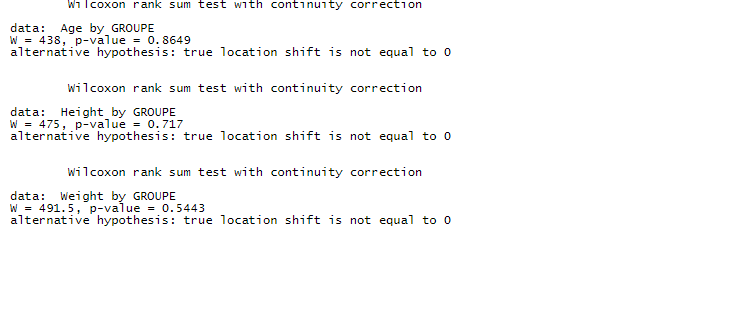


Le **test de Student** compare les moyennes des groupes NT et TSA pour chaque variable :

* **Âge** : p=0.7479p
* **Taille** : p=0.701
* **Poids** : p=0.6734p

Les p-values sont toutes supérieures à 0.05, indiquant qu'il n'y a pas de différence statistiquement significative entre NT et TSA pour ces trois variables. Cela signifie que les groupes sont bien équilibrés et comparables.

**Résultats des tests de Mann-Whitney**



Ce test non paramétrique vérifie si les distributions des groupes sont similaires, sans supposer qu'elles suivent une loi normale :

* **Âge** : p=0.8649
* **Taille** : p=0.717
* **Poids** : p=0.5443

Encore une fois, les p-values sont toutes supérieures à 0.05, confirmant qu'il n’existe pas de différence significative entre NT et TSA.

L’ensemble des résultats indique que :

* Les groupes **ne présentent pas de différences significatives** sur les variables d’appariement.
* L’appariement a été **réalisé correctement** et **ne crée pas de biais dans l’analyse des résultats**.

Nous pouvons donc conclure que l’appariement entre NT et TSA est valide et statistiquement justifié.

## Q2.2 Détection d’anomalies

Dans le cadre de l'étude de l'initiation à la marche, il est essentiel de s'assurer de la qualité des données recueillies. Une mauvaise qualité des données peut fausser les conclusions et affecter la fiabilité des résultats. Ainsi, nous avons procédé à une analyse rigoureuse des données pour identifier et corriger les valeurs manquantes et les valeurs aberrantes.

Les valeurs aberrantes ont été identifiées à l'aide de la méthode des interquartiles (IQR), où toute valeur située en dehors de l'intervalle [Q1 - 1.5\*IQR, Q3 + 1.5\*IQR] est considérée comme aberrante. Nous avons donc décidé de remplacer les valeurs aberrantes par la médiane des autres valeurs car nous trouvons ce choix judicieux et aussi ce choix aura juste un impact minimal sur notre jeu de données.

L’analyse exploratoire des données avec R a révélé la présence de valeurs manquantes dans plusieurs variables, notamment celles liées aux caractéristiques comportementales et aux indices d’évaluation cognitive. Le diagramme en barre ci-dessous nous montre bien quelles variables sont concernées.

Une image contenant texte, capture d’écran, ligne, Parallèle

Le contenu généré par l’IA peut être incorrect.

Pour les variables avec peu de valeurs manquantes, nous avons procédé à une imputation par la médiane, une approche qui préserve la distribution initiale des données.

Pour les variables fortement incomplètes, nous avons choisi de les exclure de l’analyse afin d’éviter tout biais potentiel dans l’interprétation des résultats.

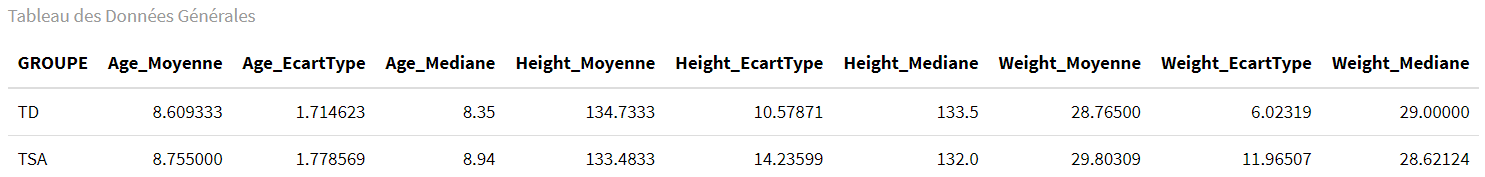
## Q2.3 : Synthèse des données descriptives

Afin de mieux comprendre les caractéristiques des participants des deux groupes (TSA et TD), nous avons procédé à une analyse descriptive des principales variables de l’étude. Cette étape exploratoire permet de dégager une première vue d’ensemble des distributions et de détecter d’éventuelles valeurs atypiques. Pour ce faire, plusieurs représentations ont été mobilisées : tableaux statistiques, boxplots et histogrammes.

### 1. Tableaux statistiques descriptifs

Afin de décrire les caractéristiques des enfants des groupes NT et TSA, nous avons réalisé une synthèse statistique sous forme de trois tableaux descriptifs. Les variables ont été regroupées selon trois catégories principales : les données générales, les données motrices et les scores relatifs aux troubles du spectre de l'autisme (TSA).

Pour chaque variable numérique sélectionnée, nous avons calculé les moyennes, écarts-types et médianes, ce qui permet de disposer à la fois d’indicateurs de tendance centrale et de dispersion. Ces mesures sont classiques en statistique descriptive et permettent de résumer efficacement les distributions observées.

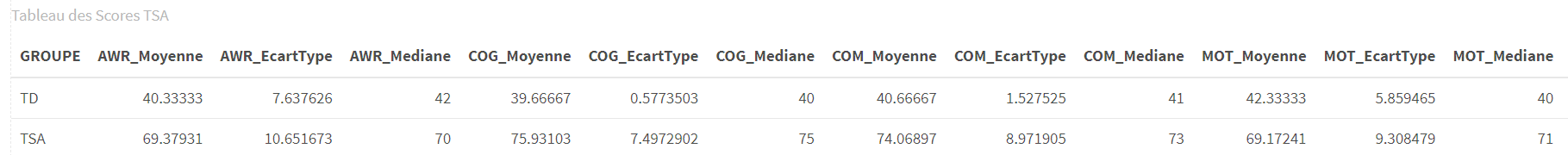


Ce tableau regroupe les données générales : âge, taille (Height) et poids (Weight).

Une image contenant texte, Police, ligne, capture d’écran

Le contenu généré par l’IA peut être incorrect.

Ce **deuxième tableau** concerne les données motrices, telles que les temps de latence (*TAju*, *TLoad*, *TUnload*) et les mesures de stabilité posturale (*CoP*).



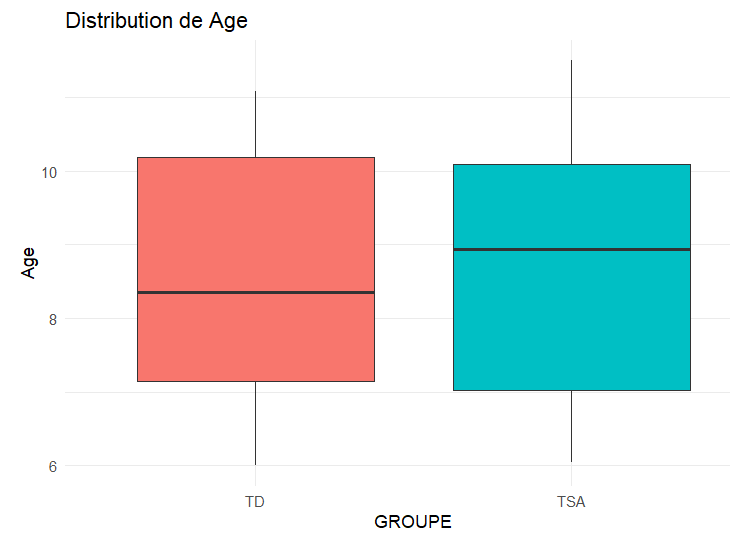
Finalement, le troisième tableau rassemble les scores issus de différentes dimensions évaluées chez les enfants TSA : communication (*COM*), motricité (*MOT*), comportement répétitif (*RRB*), etc.

Ces synthèses tabulaires permettent d’avoir une vue d’ensemble des données avant toute comparaison statistique ou modélisation.

### 2. Visualisation par boxplots

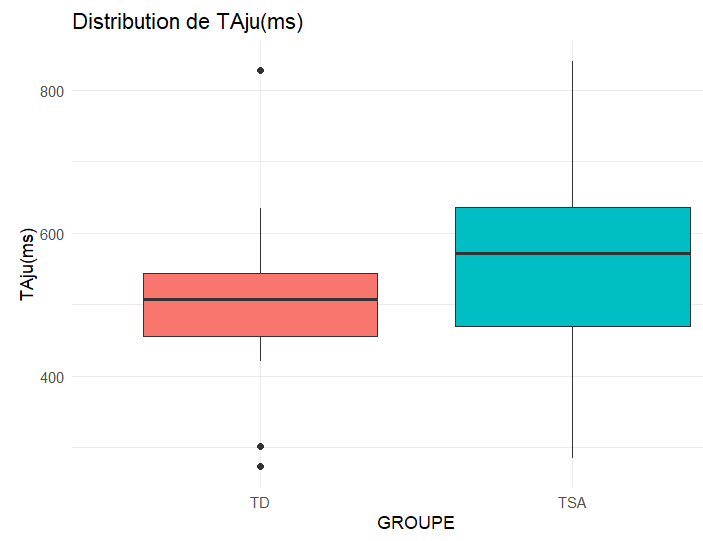
Pour compléter les tableaux, nous avons réalisé des **boxplots** afin de visualiser la distribution des principales variables par groupe. Ce type de graphique permet de :

* Visualiser la **médiane**, les **quartiles**, et les **valeurs extrêmes** (outliers)
* Comparer rapidement les **distributions** entre les groupes TSA et TD
* Repérer des **asymétries** ou des **écarts importants** de dispersion

 Une image contenant texte, capture d’écran, ligne, diagramme

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, ligne, diagramme

Le contenu généré par l’IA peut être incorrect. 

Une image contenant texte, capture d’écran, diagramme, ligne

Le contenu généré par l’IA peut être incorrect. Une image contenant texte, capture d’écran, diagramme, ligne

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, diagramme, ligne

Le contenu généré par l’IA peut être incorrect.

### Visualisation par histogrammes

Enfin, nous avons construit des **histogrammes** pour visualiser plus finement la distribution des valeurs au sein de chaque groupe. Contrairement aux boxplots, les histogrammes donnent une vision plus précise du **profil de distribution** (symétrie, concentration, étalement, etc.).

Une image contenant texte, diagramme, capture d’écran, Tracé

Le contenu généré par l’IA peut être incorrect. Une image contenant texte, diagramme, capture d’écran, Tracé

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect. Une image contenant texte, diagramme, capture d’écran, Tracé

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, diagramme, capture d’écran, Tracé

Le contenu généré par l’IA peut être incorrect. Une image contenant texte, diagramme, capture d’écran, Tracé

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, diagramme, capture d’écran, Tracé

Le contenu généré par l’IA peut être incorrect.

## Q3.1 – Vérification des conditions d’application des tests statistiques

L’objectif de cette étape est de comparer les enfants des groupes TD et TSA sur un ensemble de variables quantitatives (âge, taille, poids, paramètres biomécaniques) à l’aide de tests statistiques appropriés. Afin de garantir la validité de ces tests, les conditions suivantes ont été vérifiées au préalable :

* La **normalité de distribution** des variables,
* L’homogénéité **des variances** entre les groupes.

**Vérification de la normalité**

Nous avons testé la normalité de chaque variable à l’aide du **test de Shapiro-Wilk** une méthode statistique adaptée aux petits échantillons. Ce test permet de savoir si la variable suit une distribution gaussienne, ce qui conditionne l’utilisation d’un test paramétrique (comme le test t de Student) ou non paramétrique (comme le test de Wilcoxon).

Les variables testées sont : Age, Height, Weight, TAju(ms), TLoad(ms) et TUnload(ms). Les résultats obtenus ont guidé le choix des tests dans la phase suivante.

**Vérification de l’homogénéité des variances**

L’égalité des variances entre les groupes a été évaluée avec le **test de Levene**. Ce test permet de savoir si les groupes présentent des dispersions similaires, ce qui est une condition nécessaire pour appliquer un test t de Student classique. Dans notre cas, les variances étaient jugées homogènes pour les variables concernées.

**Choix des tests comparatifs**

Le choix du test statistique a été réalisé selon les critères suivants :

* Lorsque la variable suivait une distribution normale et que les variances étaient homogènes, un **test t de Student** a été utilisé.
* Dans le cas contraire (distribution non normale ou variances inégales), un **test de Mann-Whitney-Wilcoxon**, qui est non paramétrique, a été préféré.

Ces tests ont permis de comparer les groupes TD et TSA pour chaque variable de manière indépendante.

### Présentation des résultats

Les résultats des tests sont synthétisés dans le tableau ci-dessous. Celui-ci indique, pour chaque variable, le test utilisé, la valeur p obtenue, ainsi que si la différence entre les groupes est jugée statistiquement significative (seuil de 5 %).

**Tableau – Résultats des tests statistiques entre TD et TSA**

Une image contenant texte, capture d’écran, Police, nombre

Le contenu généré par l’IA peut être incorrect.

Les tests ont été appliqués de manière systématique selon les conditions précédemment établies.

### Comparaison graphique des distributions

Afin de compléter les analyses statistiques réalisées sur les variables quantitatives, des **boxplots** ont été construits pour chacune des variables comparées entre les groupes TSA et TD.

Ce type de représentation graphique a été choisi car il permet de :

* visualiser la **médiane** de chaque groupe pour chaque variable,
* observer l’**étendue interquartile** (dispersion des 50 % centraux),
* repérer d’éventuelles **valeurs extrêmes** (outliers),
* **comparer visuellement** les distributions des groupes sur un même axe.

Les boxplots ont été réalisés pour les variables suivantes :

* Variables générales : Age, Height, Weight
* Variables motrices : TAju(ms), TLoad(ms), TUnload(ms)

Une image contenant capture d’écran, texte, Rectangle, diagramme

Le contenu généré par l’IA peut être incorrect. Une image contenant texte, capture d’écran, ligne, diagramme

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, diagramme, ligne

Le contenu généré par l’IA peut être incorrect. Une image contenant texte, capture d’écran, ligne, diagramme

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect. Une image contenant texte, capture d’écran, ligne, diagramme

Le contenu généré par l’IA peut être incorrect.

Chaque graphique représente les données des deux groupes côte à côte, avec une couleur distincte par groupe pour améliorer la lisibilité. Ces visualisations viennent **illustrer** les résultats des tests précédents. Cette approche graphique renforce l’analyse exploratoire et permet de vérifier visuellement la pertinence des tests choisis. Elle constitue une étape complémentaire dans l’évaluation de la distribution des données et des éventuelles différences entre les groupes.

## Q3.2 — Vérification d’une corrélation entre troubles de l’anticipation motrice et scores d’évaluation du TSA (SRS et ADOS)

### Méthodologie

Pour évaluer les liens potentiels entre les variables biomécaniques et les scores d’évaluation du TSA, nous avons réalisé des tests de corrélation. Deux types de corrélation peuvent être appliqués selon la distribution des variables :

* La **corrélation de Pearson** est utilisée lorsque les deux variables suivent une distribution normale. Elle permet d'évaluer une corrélation linéaire.
* La **corrélation de Spearman** est préférable si au moins une des variables ne suit pas une distribution normale. Elle évalue une association monotone, sans supposer de linéarité.

### Vérification de la normalité

Un test de normalité (Shapiro-Wilk) a été appliqué sur l’ensemble des variables biomécaniques (TAju, TLoad, TUnload, CoP\_tot\_AP, CoP\_tot\_ML) ainsi que sur les scores TSA (AWR, COG, COM, MOT, RRB, SCI, TOT). Les résultats (tableau ci-dessous) montrent que plusieurs variables ne suivent pas une distribution normale. Par conséquent, dans la majorité des cas, la **corrélation de Spearman** a été utilisée.

Une image contenant texte, Police, nombre, ligne

Le contenu généré par l’IA peut être incorrect.

### Résultats des corrélations

Ce tableau récapitulatif ci-dessous présente les tests réalisés pour chaque combinaison variable biomécanique / score TSA. Le test appliqué, le coefficient de corrélation, la p-value et la significativité sont indiqués.

Une image contenant texte, capture d’écran, nombre, menu

Le contenu généré par l’IA peut être incorrect.

**Corrélations significatives observées**

Cinq corrélations se sont révélées statistiquement significatives (p < 0.05) :

* **TAju(ms) – RRB** (Spearman, ρ = 0.495, p = 0.004)
* **TAju(ms) – SCI** (Spearman, ρ = 0.400, p = 0.0231)
* **TAju(ms) – TOT** (Spearman, ρ = 0.445, p = 0.0106)
* **TUnload(ms) – RRB** (Spearman, ρ = 0.414, p = 0.0186)
* **CoP\_tot\_AP(/Long) – AWR** (Pearson, r = -0.453, p = 0.0092)

Ces résultats sont illustrés par les nuages de points suivants :

Une image contenant texte, capture d’écran, diagramme, ligne

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, ligne, diagramme

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, ligne, diagramme

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, ligne, diagramme

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, ligne, Tracé, diagramme

Le contenu généré par l’IA peut être incorrect.

Ces associations, bien que significatives, ne permettent pas d'établir une relation causale. Elles ouvrent néanmoins la voie à des pistes d’interprétation et de réflexion sur le lien entre motricité et caractéristiques cliniques du TSA.

## Q3.3 – Clustering des enfants selon leurs profils

L’objectif de cette analyse est de regrouper les enfants en fonction de la similarité de leurs profils, sans connaissance préalable des groupes (TSA ou TD). Pour cela, nous avons choisi une méthode de **clustering non supervisé**, en l’occurrence le **K-means**, appliquée aux **scores SRS** (AWR, COG, COM, MOT, RRB, SCI, TOT), jugés comme les plus représentatifs de l’intensité des symptômes TSA.

**Préparation des données**

Nous avons d’abord sélectionné les variables SRS d’intérêt et conservé uniquement les lignes sans valeurs manquantes. Les données ont ensuite été **centrées et réduites**, comme requis pour le K-means et l’ACP.

**Choix du nombre de clusters**

Le nombre optimal de clusters a été déterminé à l’aide de la **méthode du coude**, qui consiste à observer la courbe du total des carrés intra-cluster (Within Sum of Squares) en fonction du nombre de clusters. Le coude de la courbe se situe à **k = 3**, ce qui indique que trois clusters distincts permettent une bonne représentation des données (voir graphique ci-dessous).

Une image contenant texte, ligne, Tracé, diagramme

Le contenu généré par l’IA peut être incorrect.

*Graphique : Méthode du coude – Choix optimal de k*

### Visualisation des clusters à l’aide de l’ACP

Une **Analyse en Composantes Principales (ACP)** a été réalisée pour projeter les individus sur les deux premières dimensions principales. Cette projection permet de visualiser la répartition des clusters dans l’espace factoriel.

Une image contenant diagramme, texte, ligne, Tracé

Le contenu généré par l’IA peut être incorrect.

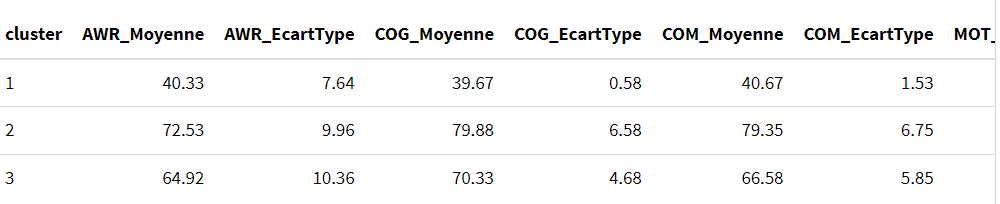
Une image contenant texte, capture d’écran, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect.

Les résultats de l’ACP montrent que la **première dimension explique à elle seule 83.9 %** de la variance totale, ce qui justifie son utilisation pour l’interprétation des clusters. Les variables les plus contributives à cette dimension sont notamment **TOT, COM, SCI, COG**, ce qui suggère que les clusters sont majoritairement différenciés sur l’intensité globale des symptômes TSA.

**Interprétation des clusters**

Un tableau récapitulatif des **moyennes et écarts-types des scores SRS** par cluster permet d’interpréter plus précisément les profils :



*Tableau : Moyennes et écarts-types des scores TSA par cluster*

## Conclusion

Ce travail a permis de mobiliser différentes méthodes statistiques pour répondre à une série de questions portant sur les différences et les similarités entre enfants neurotypiques et enfants atteints de TSA.

Après avoir vérifié la qualité des données et la validité de l’appariement entre les groupes, nous avons exploré les distributions des variables générales, motrices et comportementales.

Des tests statistiques adaptés ont été appliqués pour comparer les groupes, et des corrélations ont été recherchées afin d’évaluer les liens entre certaines variables motrices et les scores d’évaluation du TSA. Enfin, une analyse de clustering a été conduite pour identifier des regroupements d’enfants en fonction de leurs profils comportementaux.

L’ensemble des analyses a été mené de manière rigoureuse, en tenant compte des hypothèses de chaque test et en choisissant les outils graphiques les plus adaptés à chaque étape.

## Documentation

<https://documentation-snds.health-data-hub.fr/snds/fiches/appariement_donnees_externes.html>

<https://healthcaredatainstitute.com/2020/01/31/enrichissements-de-bases-de-donnees-relier-les-dossiers-cliniques-les-registres-ou-les-informations-declarees-par-les-patients-la-consommation-de-soins-de-sante/>

<https://www.qualtrics.com/fr/gestion-de-l-experience/etude-marche/analyse-cluster/>

<https://thinkr.fr/premiers-pas-en-machine-learning-avec-r-volume-3-k-means-clustering/>