PrivateTeacher

Maîtriser les Sciences Exactes

STATISTIQUES

Cours Ciblé: Statistique en Psychologie
Comparaison entre deux groupes
Test t de Student et test de Wilcoxon

Julien RUPPEN

29 March, 2026

Abstract

Ce document présente les outils statistiques disponibles pour comparer deux groupes à l'aide de données numériques : le test t de Student et le test de Wilcoxon. Il montre comment ces outils, bien que différent, suivent en réalité la même logique: construire une mesure de la différence entre deux groupes, puis quantifier la significativité statistique à l'aide de la valeur-p. Cette approche permet de couvrir un large éventail de situations différentes et d'acquérir ainsi une vue d'ensemble cohérente des tests d'hypothèse à disposition des chercheurs et chercheuses. L'approche quantitative enfin, offre un avantage important: répondre objectivement à des questions de recherche que l'observation seule ne suffit pas a clarifier.

Contents

1. Introduction : comparer deux groupes

1.1. Contexte et motivations

En psychologie clinique, on se retrouve fréquemment face à la question suivante : comment évaluer l'efficacité d'un traitement ? La réponse la plus directe à cette question consiste à comparer deux groupes identiques. L'un reçoit le traitement, l'autre non. Le groupe qui ne reçois aucun traitement sert de point de référence. On l'appel le **groupe contrôle**. On choisit ensuite une mesure que le traitement est censé modifier : un score d'anxiété (GAD-7), un niveau de dépression (BDI-II), ou encore une mesure de qualité du sommeil. Cette mesure est relevée dans chacun des deux groupes après l'intervention et les deux valeurs sont alors comparées.

La valeur de ce dispositif repose sur un principe simple : si les deux groupes sont identiques au départ, la seule différence entre les deux est le traitement. Si une différence de la mesure choisie est observé entre les groupe, celle ci est donc attribuable au traitement. Si aucune différence n'est détectée, le traitement est jugé sans effet. Le test t de Student et le test de Wilcoxon sont des instruments mathématiques qui permettent de déceler cette différence même lorsqu'elle n'est pas visible à l'œil nu — une sorte de loupe numérique appliquée aux données.

> **Point Important :** le test d'hypothèse n'est utile *que* lorsqu'il y a recouvrement entre les deux distributions. Si les deux groupes sont parfaitement séparés, la différence est évidente à l'œil nu et aucun test n'est nécessaire.

1.1.1 Représentation graphique

Dans l'exemple représenté ici, les deux courbes sont décalées — le groupe traitement présente en moyenne des scores plus élevés (38.6) que le groupe contrôle (23.7) — mais elles se superposent dans une zone intermédiaire. Cette superposition n'est pas un artefact : elle reflète la réalité clinique, où les individus varient. Le test statistique va quantifier si cet écart de 15 points dépasse ce qu'on attendrait par simple hasard d'échantillonnage.

1.1.2 Questions de recherche

Dans ce genre de situation, on se pose typiquement la question suivantes: Les deux groupes diffèrent-ils significativement sur le score mesuré ?

1.2. Méthode de résolution

Face à cette situation, le raisonnement est le suivant: On commence par **choisir une mesure de la différence** adaptée à la nature des données : le test t de Student si les données sont approximativement normales et les groupes indépendants, le test d

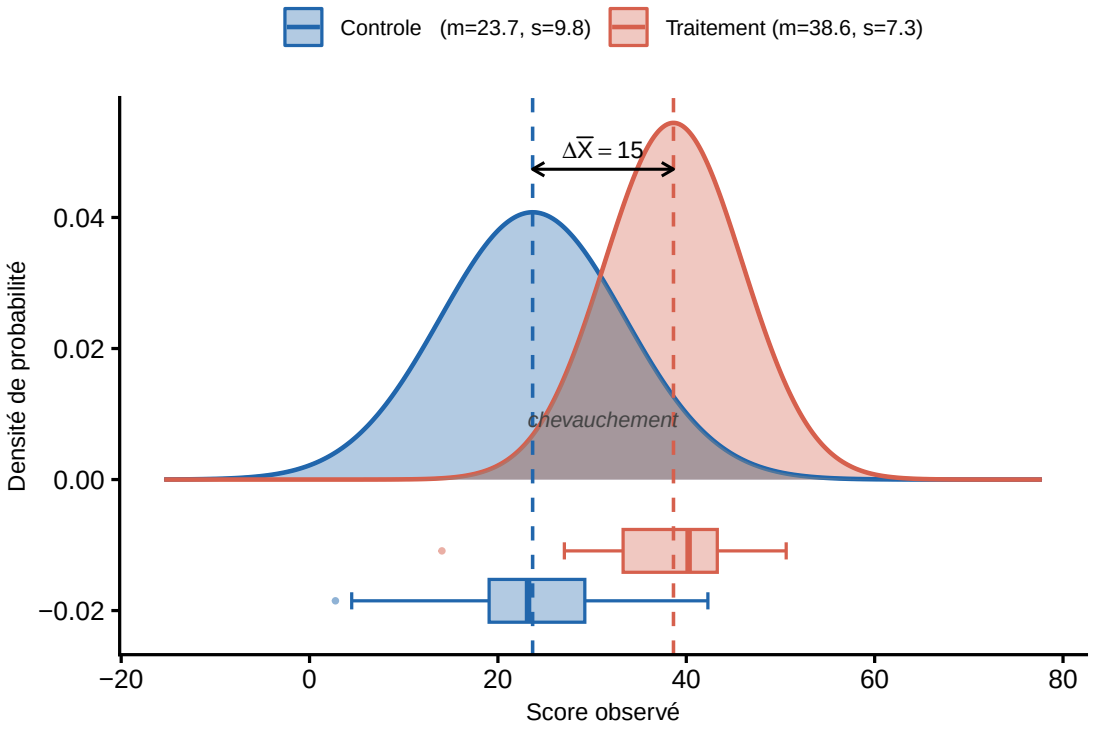


Figure 1: Distributions du score clinique dans le groupe contrôle et le groupe traitement. La zone grisée représente la région de chevauchement — là où un score seul ne permet pas de distinguer les deux groupes. Les lignes verticales indiquent les moyennes respectives ; les boxplots horizontaux résument la dispersion.

Wilcoxon si les données sont ordinales ou si la normalité est violée. On calcule ensuite cette mesure. Le test statistique produit alors une **p-valeur** qui quantifie la probabilité d'observer un écart au moins aussi grand que celui mesuré, si les deux groupes étaient en réalité identiques. Une p-valeur petite signifie que cette probabilité est faible — autrement dit, que la différence observée est difficile à expliquer par le seul hasard.

Remarque: une p-valeur significative ne dit rien de l'*importance clinique* de la différence. Elle permet uniquement de dire si la différence est réelle, c'est à dire, qu'elle n'est pas due au hasard. Une différence de 0.3 point sur une échelle de 100 peut être hautement significative, mais ne pas être cliniquement pertinente. C'est pourquoi on complète toujours le test par une **taille d'effet** (Cohen's d, corrélation rang-bisériale r), qui mesure l'ampleur de la différence indépendamment de la taille de l'échantillon.

On résumera cette logique à l'aide des trois étapes suivantes :

1. Choisir une mesure adaptée à la situation (choisir le test d'hypothèse)
2. Déterminer la significativité statistique de la mesure (evaluer la valeur p)
3. Interpréter le résultat numérique dans le context de la question de recherche.

Cette même séquence sera appliquée dans tous les exemples qui suivent.

2. Cas 1 : deux groupes indépendants

2.1. Résumé de la situation

Ici, un groupe de **250 patients** a suivi une thérapie cognitivo-comportementale (TCC), tandis qu'un groupe de **320 patients** a été placé en liste d'attente et n'ont donc pas reçu de traitement (groupe contrôle). Les deux groupes ont été évalués à l'aide du **GAD-7** (Generalized Anxiety Disorder scale, 7 items, score 0–21), l'un des instruments les plus utilisés en clinique pour mesurer l'anxiété généralisée. Un score élevé indique une anxiété plus sévère.

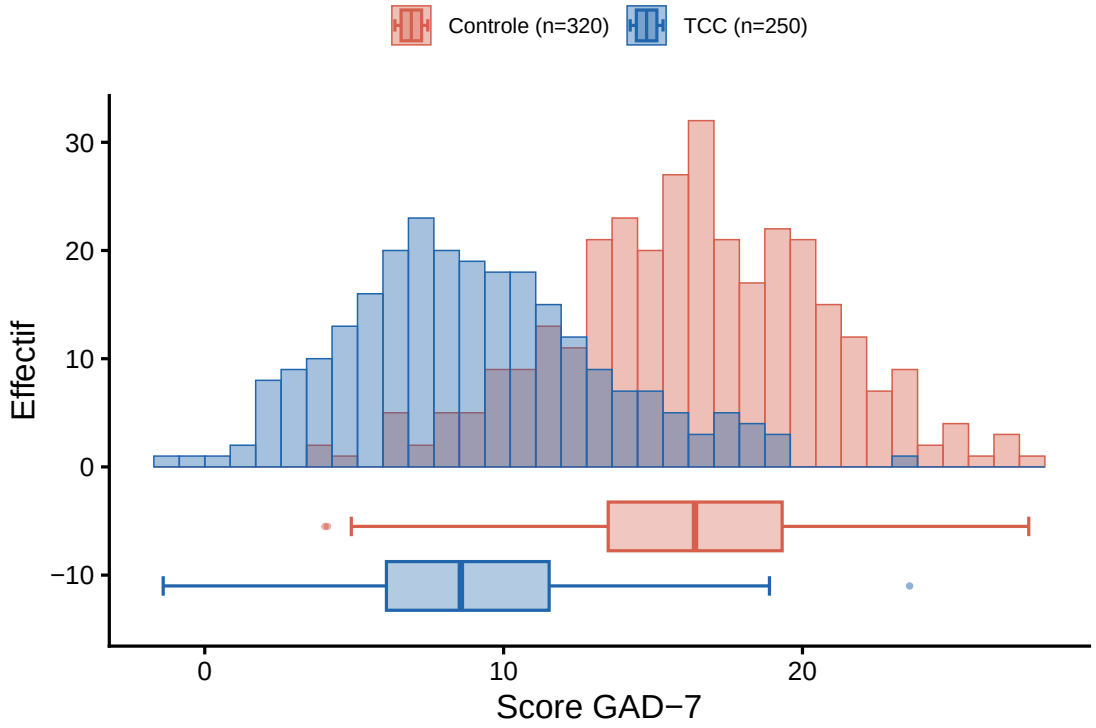


Figure 2: Distributions des scores GAD-7 dans les deux groupes. Les histogrammes superposés montrent le chevauchement partiel des distributions ; les boxplots horizontaux résument la position et la dispersion. La séparation entre les médianes est nette.

L'inspection visuelle suggère déjà une différence marquée : le groupe TCC est centré sur des scores bas, le groupe contrôle sur des scores élevés. Mais peut-on conclure que cette différence est statistiquement significative ?

2.2. Question de recherche:

Les scores GAD-7 diffèrent-ils significativement entre le groupe TCC et le groupe contrôle ?

2.3. Méthode de résolution

Étape 1 — Choisir une mesure adaptée. Les données sont continues (scores GAD-7, valeurs entières de 0 à 21), les deux groupes sont indépendants, et les effectifs sont largement suffisants (n = 250 et n = 320) pour invoquer le théorème central limite : la distribution des moyennes sera approximativement normale même si les scores individuels ne le sont pas exactement. On utilise le **test *t* de Student pour deux groupes indépendants**, avec l'hypothèse de variances égales (`var.equal = TRUE`).

> **Pourquoi `var.equal = TRUE` ?** L'argument suppose que les deux populations ont la même variance. Si ce n'est pas le cas, on utilise `var.equal = FALSE` (test de Welch), qui ajuste les degrés de liberté en conséquence. Avec des effectifs importants et similaires en ordre de grandeur, les deux variantes donnent des résultats très proches.

Étape 2 — Calculer et interpréter la p-valeur. La commande R calcule la statistique t et la p-valeur. On comparera cette p-valeur au seuil $\alpha = 0.05$.

Étape 3 — Décision. Si $p < 0.05$, on conclut que les groupes diffèrent significativement sur le GAD-7.

2.4. Commande R

```
t.test(groupe_controle, groupe_tcc, var.equal = TRUE)
```

2.5. Interprétation des Résultats

2.5.1 Sortie logicielle

```
    Two Sample t-test

data:  groupe_controle_sect2 and groupe_tcc_sect2
 t = 20.073, df = 568, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 6.615507 8.050611
sample estimates:
mean of x mean of y
16.294537  8.961478
```

2.5.2 Interprétation

La p-valeur est très petite — bien en dessous de tout seuil de 0.05. On rejette donc l'hypothèse nulle sans hésitation : les deux groupes ne sont pas équivalents sur le GAD-7.

Le groupe contrôle présente des scores nettement plus élevés (16.3) que le groupe TCC (9), soit un écart moyen de 7.3 points. La taille d'effet (Cohen's $d = 1.694$) est **très large** — un effet de cette magnitude est rare en psychologie clinique et indique que les deux groupes se distinguent nettement, pas seulement statistiquement.

> **Mise en garde sur les grands effectifs.** Avec n = 250 et n = 320, le test a une puissance statistique très élevée : même une différence cliniquement négligeable (quelques dixièmes de point) pourrait atteindre la significativité. C'est pourquoi le d de Cohen est ici indispensable : il confirme que l'écart observé est non seulement statistiquement significatif, mais aussi cliniquement substantiel.

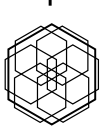
2.5.3 Réponse

> **Oui.** Une différence très hautement significative est détectée : $t(568) = 20.073$, $p < 2.2e\text{-}16$. Le groupe contrôle présente un score GAD-7 moyen de 16.3 contre 9 dans le groupe TCC ($d = 1.694$).

2.6. Vérification

Deux repères visuels confirment le résultat. D'abord, les histogrammes : les deux distributions sont clairement décalées, avec un chevauchement limité à la zone 10–15 environ. Ensuite, les boxplots : les IQR des deux groupes ne se recouvrent pas — la médiane du groupe contrôle est au-dessus du troisième quartile du groupe TCC.

L'intervalle de confiance à 95 % [6.62 ; 8.05] est entièrement positif et éloigné de zéro. Cela signifie que, quelle que soit l'incertitude d'échantillonnage, la différence réelle est très vraisemblablement comprise dans cet intervalle — et elle n'est certainement pas nulle.



3. Cas 2 : données ordinales

3.1. Résumé de la situation

Le test t repose sur une hypothèse fondamentale : les données sont des valeurs continues issues d'une distribution approximativement normale, et les *moyennes* constituent une mesure utile de la tendance centrale. Cette hypothèse cependant n'est valable que si les effectifs sont importants.

Une échelle ordinale permet de classer les individus selon une valeur numérique ($1 < 2 < 3 < \ldots < 7$), mais les intervalles entre les catégories ne sont pas nécessairement égaux. La différence entre "1 — symptômes absents" et "2 — symptômes légers" n'est pas forcément la même que la différence entre "6 — symptômes sévères" et "7 — symptômes très sévères". Calculer une moyenne sur ces valeurs et appliquer un test t reviendrait à traiter ces catégories comme si elles étaient équidistantes — ce qui n'est pas justifié.

Dans cette étude, 45 patients (groupe TCC) et 38 patients (groupe contrôle, liste d'attente) ont évalué la **sévérité de leurs symptômes anxieux** sur une échelle de 1 à 7. L'outil approprié ici est le **test de Wilcoxon–Mann–Whitney**, qui transforme les valeurs brutes en *rangs* avant de comparer les groupes — une opération qui ne suppose aucune équidistance entre les catégories.

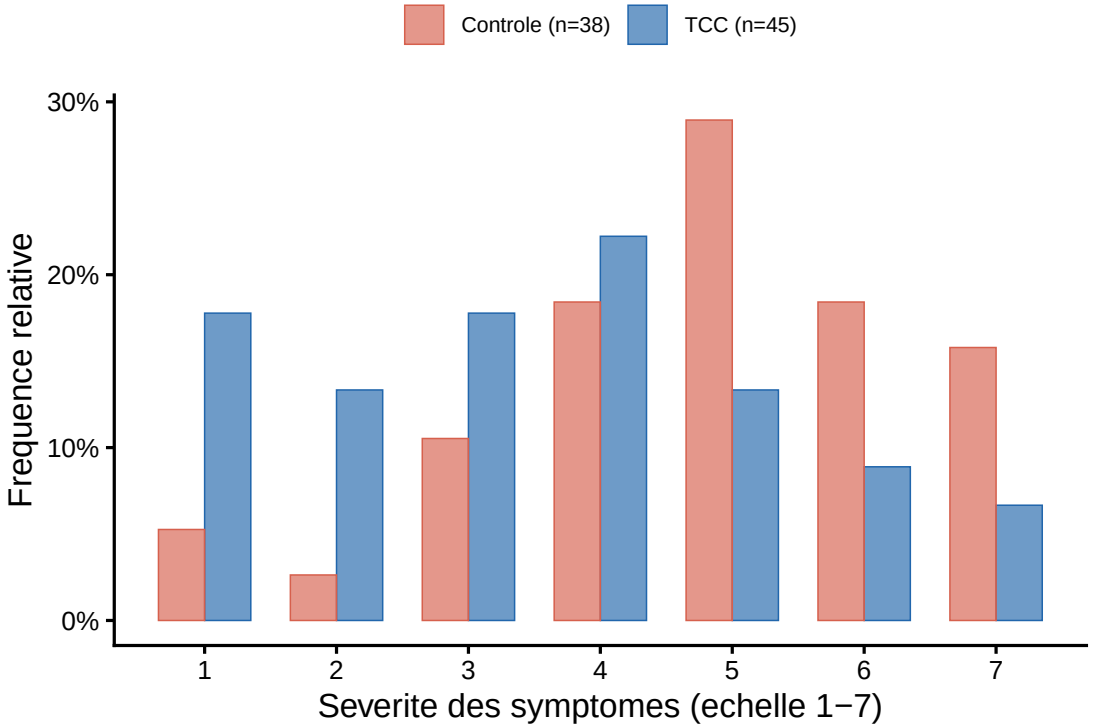


Figure 3: Fréquences relatives des scores de sévérité (1–7) dans les deux groupes. Le groupe TCC est clairement concentré sur les valeurs basses (symptômes faibles) ; le groupe contrôle sur les valeurs hautes (symptômes sévères).

On voit immédiatement que les distributions sont inversées : TCC domine les modalités 2–4, contrôle domine les modalités 4–6.

3.2. Question de recherche

Les scores de sévérité diffèrent-ils significativement entre le groupe TCC et le groupe contrôle ?

3.3. Méthode de résolution

Étape 1 — Choisir une mesure adaptée. Les données sont ordinales. Le test de Wilcoxon attribue un rang global à chaque observation (tous groupes confondus), puis compare la somme des rangs entre les deux groupes. La statistique W compte le nombre de paires (contrôle_i, TCC_j) pour lesquelle le score du groupe contrôle dépasse du score du groupe TCC. C'est l'équivalent de la statistique de test t.

Étape 2 — Calculer et interpréter. Une fois la valeur de W calculé, on cherche a savoir si cette valeur est importante ou non. Plus W s'éloigne de la valeur centrale, plus la différence entre les groupes est marquée. Tout comme dans le cas du test t, on utilise alors la p-valeur pour quantifier la significativité statistique de l'écart et pour statuer si cet écart est dû au hasard ou s'il est dû au traitement.

Étape 3 — Décision. Si $p < 0.05$, les deux distributions sont différente, si $p > 0.05$ alors les deux distribution ne sont pas significativement différente et il n'esta pas possible de dire que le traitement à eu un effet.

> **Note :** pour les données ordinales, on rapporte la **médiane** (et non la moyenne) comme mesure de tendance centrale. La médiane contrôle est 5, la médiane TCC est 4.

3.4. Commande R

```
wilcox.test(groupe_controle, groupe_tcc, exact = FALSE)
```

3.5. Interprétation des Résultats

3.5.1 Sortie logicielle

```
    Wilcoxon rank sum test with continuity correction

data:  groupe_controle_sect3 and groupe_tcc_sect3
 W = 1206, p-value = 0.00117
alternative hypothesis: true location shift is not equal to 0
```

3.5.2 Interprétation

La p-valeur (0.00117) est un ordre de grandeur inférieure au seuil $\alpha = 0.05$: la différence est donc statistiquement significative. La statistique $W = 1206$ est nettement supérieure à la valeur centrale attendue sous H0 (855), confirmant que les scores contrôle tendent à dépasser les scores TCC dans la majorité des paires.

La médiane contrôle (5) est supérieure à la médiane TCC (4). La corrélation rang-bisériale $|r| = 0.411$ indique un effet de taille **modérée** — perceptible sur le barplot, mais moins marqué que dans le cas précédent. Cela illustre une situation clinique fréquente : la TCC produit un effet réel, mais la variabilité individuelle sur une échelle à 7 points limite la magnitude de l'effet détectable.

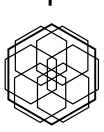
3.5.3 Réponse

> **Oui.** Une différence significative est détectée : $W = 1206$, $p = 0.00117$. Le groupe contrôle présente une sévérité médiane de 5 contre 4 dans le groupe TCC ($|r| = 0.411$).

3.6. Vérification

Le barplot offre ici une vérification intuitive directe. Pour chaque modalité, comparez les deux barres : aux valeurs 1–3, la barre TCC domine ; aux valeurs 5–7, la barre contrôle domine. Cette inversion systématique est exactement ce que le test de Wilcoxon détecte — un décalage des rangs. Si les deux groupes étaient identiques, les barres seraient à peu près de même hauteur pour chaque modalité.

Un repère numérique : $W = 1206$ contre une valeur centrale de 855 sous H0. L'écart est de 351 unités — soit environ 41% de déviation par rapport à la valeur centrale. Cela confirme un déséquilibre réel entre les rangs des deux groupes.



4. Cas 3 : mesure appariée (pré/post)

4.1. Résumé de la situation

Dans les deux cas précédents, les participants des deux groupes étaient des personnes *différentes* — on parle alors de groupes **indépendants**. Il existe une configuration tout aussi fréquente en psychologie clinique: les **mêmes individus** sont mesurés à deux reprises, avant et après une intervention. On parle alors de **mesure appariée** ou de plan pré/post.

L'avantage de cette structure est important : en comparant chaque individu à lui-même, **on élimine la variabilité** qui tient aux différences entre personnes (personnalité, niveau de base, tolérance au stress). Ce qu'on mesure réellement, c'est le *changement individuel* — et le signal devient beaucoup plus net par rapport au bruit. À effectif égal, un plan apparié détectera des effets plus petits qu'un plan indépendant.

Dans cette étude, 28 participants ont complété le **PSS-10** (Perceived Stress Scale, 10 items, score 0–40) avant et après un programme de pleine conscience de huit semaines. Un score élevé indique un stress perçu plus intense. L'objectif est de déterminer si le programme a significativement réduit le stress.

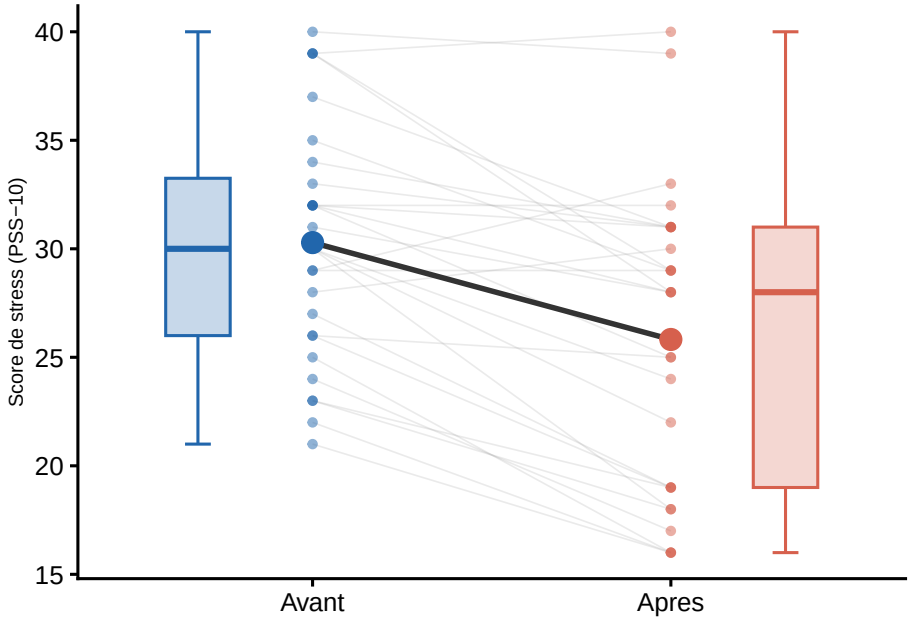


Figure 4: Score de stress (PSS-10) pour chaque participant avant (Avant) et après (Apres) l'intervention. Chaque ligne grise représente la trajectoire individuelle d'un participant. La ligne noire relie les moyennes des deux temps de mesure.

Ce type de graphique — un *spaghetti plot* — révèle deux informations complémentaires. La pente de la ligne noire (moyennes) indique la direction et l'ampleur de l'effet moyen. Les lignes individuelles révèlent la variabilité des trajectoires : la majorité descend (amélioration), mais quelques participants voient leur score augmenter. Cette variabilité individuelle est normale et attendue — aucune intervention n'est efficace pour 100 % des participants.

4.2. Question de recherche

Le niveau de stress des participants a-t-il significativement diminué après l'intervention ?

4.3. Méthode de résolution

Étape 1 — Choisir une mesure adaptée. Les observations sont couplée (on dit aussi: appariées (même individu mesuré deux fois)). La bonne approche consiste donc à calculer la **différence individuelle** (post - pré) pour chaque participant, puis à tester si la moyenne des différences différente de zéro. Ce changement de perspective est un point clé : plutôt que de mesurer la différence des moyenne (groupe indépendant) on mesure à présent **la moyenne des différences** de chaque individu. En procédant ainsi, la moyenne des différences tient compte de la corrélation entre les deux mesures, ce qui réduit la variabilité entre les individu et augmente la puissance du test. C'est une force du test t apparié.

Étape 2 — Calculer et interpréter. La sortie R affichera la différence moyenne entre post et pré (-4.46 points ici), la statistique t et la p-valeur associée. La taille d'effet pour données appariées est Cohen's d = moyenne des différences / écart-type des différences.

Étape 3 — Décision. Si $p < 0.05$, la réduction moyenne du score PSS-10 est statistiquement significative.

4.4. Commande R

```
t.test(post, pre, paired = TRUE)
```

4.5. Interprétation des Résultats

4.5.1 Sortie logicielle

```
    Paired t-test

data:  post_sect4 and pre_sect4
 t = -5.8855, df = 27, p-value = 2.862e-06
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -6.020636 -2.907935
sample estimates:
mean difference
```

```
-4.464286
```

4.5.2 Interprétation

La p-valeur (2.86e-06) est très inférieure au seuil $\alpha = 0.05$: la réduction du stress est statistiquement significative. En moyenne, les participants ont diminué leur score PSS-10 de 4.46 points (de 30.3 à 25.8).

La taille d'effet est large (Cohen's $d = 1.112$), ce qui indique que la réduction est non seulement réelle mais cliniquement substantielle — d'autant plus remarquable compte tenu du petit effectif (n = 28). C'est précisément l'avantage du plan apparié : avec un petit échantillon bien contrôlé, on peut détecter des effets qui nécessiteraient des centaines de participants dans un plan indépendant.

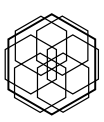
4.5.3 Réponse

> **Oui.** La réduction du stress est hautement significative : $t(27) = -5.886$, $p =$ 2.86e-06. La différence moyenne est de -4.46 points ($d = 1.112$).

4.6. Vérification

Deux repères permettent de valider le résultat. Sur le graphique, comptez les lignes qui descendent versus celles qui montent : la majorité descend nettement, ce qui traduit une amélioration répandue. La ligne noire des moyennes confirme la direction et l'ampleur de cet effet moyen.

Numériquement, l'intervalle de confiance à 95 % [-6.02 ; -2.91] est entièrement négatif et éloigné de zéro. Cela signifie qu'on est très confiant que la vraie réduction moyenne se situe quelque part dans cet intervalle — et qu'elle est certainement différente de zéro. Un intervalle qui chevaucherait zéro signalerait au contraire une incertitude sur la direction même de l'effet.



5. Conclusion

5.1. Ce que nous avons appris

Dans ce document, nous avons vu comment plusieurs tests d'hypothèse s'organisent tous autour de la même logique:

1. **Choisir une mesure adaptée** — choisir le bon test d'hypothèse.
2. **Déterminer la significativé statistique de la mesure** — évaluer la p-valeur.
3. **Prendre une décision** — interpréter la p-valeur et répondre à la question.

Nous l'avons vu dans chacun de nos exemples, cette logique se retrouve dans toutes les situations où l'on cherche à comparer deux groupes sur la base de données numériques. La différence entre les tests quant à elle apparaît lorsque la nature des variables change (continues ou ordinales) ou lorsque les prérequis de normalité ou d'homogénéité des variances ne sont pas vérifiés.

5.2. Méthode : choisir le bon test

Le tableau suivant résume ces critères :

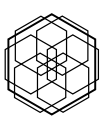
Situation	**Test**	**Fonction R**
Données continues, variances égales, groupes indépendants	Test t de Student	`t.test(..., var.equal=TRUE)`
Données continues, variances inégales, groupes indépendants	Test de Welch	`t.test(..., var.equal=FALSE)`
Données ordinales ou normalité violée, groupes indépendants	Test de Wilcoxon	`wilcox.test(...)`
Mêmes individus mesurés deux fois (design pré/post)	Test t apparié	`t.test(..., paired=TRUE)`

5.3. Travailler avec méthode

Dans ce document, nous avons vu comment une même séquence d'étapes s'applique à plusieurs cas de figure différents. Cette séquence permet de construire un raisonnement que l'on peut utiliser dans de nombreuses situations, ce qui simplifie considérablement le raisonnement. Il est dès lors possible de développer une intuition sur la manière d'utiliser les outils à notre disposition et d'interpréter les résultats dans le contexte particulier de la question de recherche qui nous intéresse.

5.4. Prendre des points de repère

Il convient enfin de faire attention avec la manipulation de données numériques car un ordinateur calcule facilement nimporte quelle valeurque les données soit correctes ou qu'elle soit erronnée. Un ordinateur en effet **ne détecte pas les erreurs de saisie**, les



mauvaises variables ou les groupes confondus. Pour cette raison, voici comment vérifier ses résultats à l'aide de trois repères simples à mettre en place systématiquement

1. Le graphique. vérifier visuellement que le résultat numérique est cohérent avec le graphique. considérer les valeur p valeur de deux cas extrêmes pour se donner un point de repère: recouvrement total → p-valeur proche de 1 ; séparation totale → p-valeur proche de 0.

2. Les moyennes. Pour les tests t, la sortie R affiche les moyennes des deux groupes. Vérifiez que ces valeurs correspondent aux valeur calculée manuellement sur les données brutes. C'est un moyen rapide de savoir si l'on travail avec les bonnes données.

3. L'intervalle de confiance à 95% Si cette intervalle contient la valeur zéro, cela confirm que la différence n'est probablement pas significative. Cela doit correspondre avec la conclusion tirée de l'interprétation de la valeur p. Les deux interprétation doivent conduire au meme conclusion, signe que l'interprétation est correct.