

Estimation de la condition physique liée à la santé

par régression linéaire multiple chez les adultes coréens

National Fitness Award 2015–2019

Auteurs

Farah Ben Yedder

Wael Ben Salem

Sirine Hjaiej

Mohamed Abdennadher

Zakaria Bouchaddakh

Aziz Hamlaoui

**École Supérieure Privée d'Ingénierie et
de Technologies (ESPRIT)**
Département DATA SCIENCE

15 décembre 2025

Introduction

Contexte général

La condition physique liée à la santé (Health-Related Physical Fitness, HRPF) est un indicateur essentiel de l'état de santé global. Elle regroupe plusieurs composantes clés, notamment la force musculaire, l'endurance, la flexibilité et la capacité cardiorespiratoire. Un niveau adéquat de condition physique est reconnu comme un facteur majeur de prévention des maladies chroniques telles que les pathologies cardiovasculaires, le diabète de type 2 et l'obésité.

Cependant, l'évaluation directe de ces paramètres repose souvent sur des tests en laboratoire coûteux, chronophages et peu accessibles à grande échelle, ce qui limite leur utilisation dans les études populationnelles et le suivi préventif.

Problématique et objectifs

Afin de surmonter ces contraintes, il est nécessaire de développer des méthodes de prédiction simples et fiables permettant d'estimer la condition physique à partir de variables facilement mesurables. La présente

étude a pour objectif de développer et valider des modèles de régression linéaire multiple permettant de prédire plusieurs paramètres de condition physique à partir de variables anthropométriques et démographiques.

Les objectifs spécifiques sont :

- Décrire les caractéristiques de l'échantillon étudié
- Explorer les relations entre variables anthropométriques et performances physiques
- Évaluer la distribution des variables et sélectionner les tests statistiques appropriés
- Comparer les performances physiques selon le sexe
- Développer et valider des équations de prédiction des paramètres HRPF

Données et méthodologie

Les données analysées proviennent du **National Fitness Award (NFA)** de Corée du Sud, collectées entre 2015 et 2019. L'échantillon comprend **2 000 adultes âgés de 19 à 64 ans**. L'analyse repose sur une démarche progressive intégrant la préparation des données, l'analyse descriptive, les tests de normalité, les tests d'hypothèses ainsi que la modélisation par régression linéaire multiple.

Structure du rapport

Le rapport est structuré en cinq parties principales :

- 1. Préparation des données
- 2. Analyse descriptive
- 3. Tests sur la normalité
- 4. Tests d'hypothèses paramétriques et non-paramétriques
- 5. Modélisation par Régression Linéaire Multiples

PHASE 1 : PRÉPARATION DES DONNÉES

1. Importation des données et nettoyage des noms

Cette première étape consiste à importer le jeu de données brut, charger les bibliothèques nécessaires à l'analyse statistique et harmoniser les noms des variables afin de garantir une manipulation cohérente et reproductible des données.

Bibliothèques utilisées et leur utilité

- `lmtest` : permet de réaliser des tests statistiques sur les modèles de régression, notamment les tests d'hétéroscédasticité et d'autocorrélation.
- `broom` : facilite l'extraction et la mise en forme des résultats des modèles statistiques sous forme de tableaux exploitables.
- `tidyverse` : regroupe plusieurs packages essentiels pour la manipulation, la visualisation et l'analyse cohérente des données.
- `car` : fournit des outils avancés pour le diagnostic et l'évaluation des modèles de régression.
- `naniar` : permet l'exploration et la visualisation des données manquantes.
- `readxl` : permet l'importation de fichiers Excel dans l'environnement R.
- `janitor` : sert à nettoyer et harmoniser les noms des variables afin d'améliorer la lisibilité et la reproductibilité de l'analyse.
- `lubridate` : facilite la manipulation et le traitement des variables de type date et heure.
- `VIM` : offre des méthodes de visualisation et d'imputation des valeurs manquantes.
- `GGally` : étend les fonctionnalités de `ggplot2`, notamment pour la création de matrices de corrélation.
- `tibble` : propose une version moderne et améliorée des data frames pour une meilleure lisibilité.
- `kableExtra` : permet de produire des tableaux esthétiques et professionnels pour les rapports statistiques.
- `dplyr` : facilite la manipulation des données à travers des opérations de filtrage, sélection et transformation.
- `psych` : fournit des outils pour l'analyse descriptive et psychométrique des données.
- `normtest` : permet de réaliser des tests statistiques de normalité.
- `knitr` : assure l'intégration du code R et de ses résultats dans des documents reproductibles.
- `ggplot2` : permet la création de graphiques statistiques clairs, cohérents et personnalisables.

```
library(lmtest)
library(broom)
library(tidyverse)
library(car)
library(naniar)
library(readxl)
library(janitor)
```

```

library(lubridate)
library(VIM)
library(GGally)
library(tibble)
library(kableExtra)
library(dplyr)
library(psych)
library(nortest)
library(knitr)
library(ggplot2)

```

Importation du dataset initial :

```

df_initial <- read_excel(
  "C:/Users/ASUS/Downloads/stat/Projet Stat/DATA.xlsx",
  na = ""
)

```

Nettoyage des noms de colonnes :

```

df_initial <- df_initial %>%
  clean_names()

```

Aperçu général :

```
glimpse(df_initial)
```

```

## Rows: 2,000
## Columns: 10
## $ participant_id      <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, ~
## $ sex                  <chr> "M", "M", "M", "F", "F", "M", "M"~
## $ age                  <dbl> 46, 25, 53, 46, 32, 28, 34, 34, 49, 60, 43, ~
## $ measurement_date     <chr> "2019-01-19", "2018-03-16", "2017-06-02", "~"
## $ bmi                  <dbl> 26.5, 28.8, 25.7, 28.0, 25.0, 24.7, 24.7, 2~
## $ percent_body_fat     <dbl> 45.0, 45.0, 45.0, 45.0, 57.7, 60.0, 60.0, 4~
## $ hand_grip_strength_kg <dbl> 47.0, 28.0, 37.7, 48.9, 20.0, 21.2, 28.9, 3~
## $ sit_and_reach_cm      <dbl> 13.4, -1.2, 4.5, 9.0, 14.8, 13.1, 3.9, 17.1~
## $ sit_ups_count         <dbl> 13, 19, 18, 20, 23, 27, 17, 12, 12, 15, 16, ~
## $ vo2_estimate_ml_per_kg_min <dbl> 29.8, 46.5, 37.2, 52.5, 24.6, 30.1, 30.0, 4~

```

```
summary(df_initial)
```

```

##   participant_id       sex           age   measurement_date
##   Min. : 1.0   Length:2000   Min. :19.00   Length:2000
##   1st Qu.: 500.8  Class :character  1st Qu.:30.00  Class :character
##   Median :1000.5  Mode  :character  Median :41.00  Mode  :character
##   Mean   :1000.5                   Mean   :41.54
##   3rd Qu.:1500.2                   3rd Qu.:53.00
##   Max.   :2000.0                   Max.   :64.00
##   bmi            percent_body_fat hand_grip_strength_kg sit_and_reach_cm
##   Min.   :15.00    Min.   :26.50    Min.   : 7.40    Min.   :-14.900

```

```

##   1st Qu.:21.80    1st Qu.:45.00    1st Qu.:27.88    1st Qu.: 3.700
##   Median :24.00    Median :45.00    Median :36.40    Median : 7.900
##   Mean    :23.98    Mean   :47.80    Mean   :35.96    Mean   : 7.887
##   3rd Qu.:26.10    3rd Qu.:53.33    3rd Qu.:43.70    3rd Qu.:11.900
##   Max.    :37.10    Max.   :60.00    Max.   :68.00    Max.   :34.400
##   sit_ups_count    vo2_estimate_ml_per_kg_min
##   Min.    : 5.00    Min.   :10.00
##   1st Qu.:15.00    1st Qu.:31.50
##   Median :17.00    Median :36.40
##   Mean   :17.62    Mean   :36.28
##   3rd Qu.:20.00    3rd Qu.:41.20
##   Max.   :33.00    Max.   :60.10

```

```
df_initial %>% tabyl(sex)
```

```

##   sex      n percent
##   F     645  0.3225
##   M    1355  0.6775

```

Interprétation : Le jeu de données comprend 2 000 adultes âgés de 19 à 64 ans, avec une majorité d'hommes. Les variables anthropométriques et de performance physique présentent une variabilité importante, ce qui rend l'échantillon adapté aux analyses statistiques et aux modèles de régression.

2. Conversion des types et extraction temporelle

Les variables sont ensuite converties vers des types adaptés aux analyses statistiques.

Une extraction de l'année et du mois à partir de la date de mesure est réalisée afin d'introduire une dimension temporelle.

```

df_initial <- df_initial %>%
mutate(
  participant_id = as.integer(participant_id),
  sex = as.factor(sex),
  age = as.integer(age),
  measurement_date = as.Date(measurement_date),
  bmi = as.numeric(bmi),
  percent_body_fat = as.numeric(percent_body_fat),
  hand_grip_strength_kg = as.numeric(hand_grip_strength_kg),
  sit_and_reach_cm = as.numeric(sit_and_reach_cm),
  sit_ups_count = as.integer(sit_ups_count),
  vo2_estimate_ml_per_kg_min = as.numeric(vo2_estimate_ml_per_kg_min),
  measurement_year = year(measurement_date),
  measurement_month = month(measurement_date)
)

```

Interprétation : La conversion des types garantit la validité des analyses statistiques. L'extraction de l'année et du mois permet d'explorer d'éventuelles variations temporelles ou saisonnières des performances physiques.

3. Vérification des valeurs manquantes

```
# Vérification des valeurs manquantes
sum(is.na(df_initial))
```

```
## [1] 0
```

Interprétation : Aucune valeur manquante détectée (0 NA) → la base de données est complète pour toutes les variables et tous les participants, permettant d'effectuer les analyses sans traitement préalable des données manquantes ni réduction de l'échantillon.

4. Analyse des valeurs aberrantes

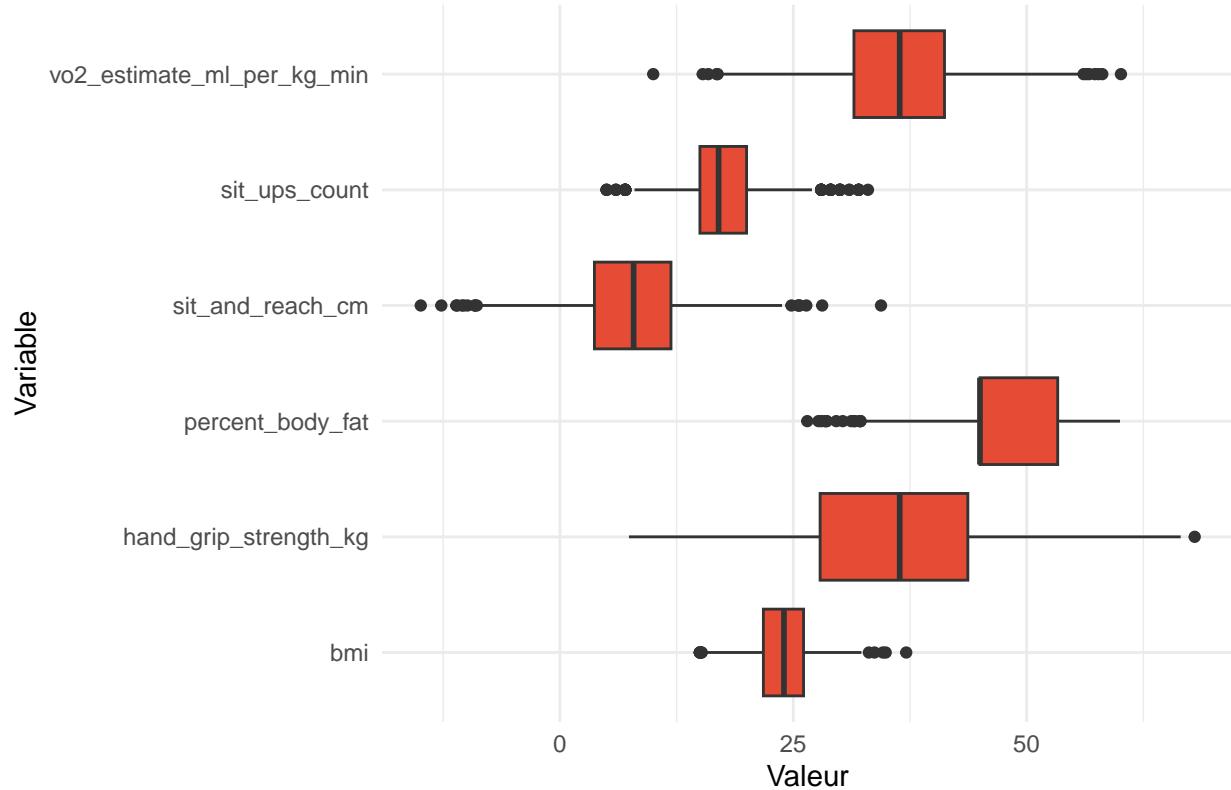
La présence de valeurs aberrantes est évaluée afin de vérifier l'intégrité du jeu de données.

```
vars_to_exclude <- c(
  "participant_id",
  "age",
  "measurement_year",
  "measurement_month"
)

numeric_vars <- df_initial %>%
  select(where(is.numeric)) %>%
  select(-all_of(vars_to_exclude))

numeric_vars %>%
  pivot_longer(everything()) %>%
  ggplot(aes(x = name, y = value)) +
  geom_boxplot(fill = "#E64B35") +
  coord_flip() +
  theme_minimal() +
  labs(
    title = "Boxplots des variables numériques",
    x = "Variable",
    y = "Valeur"
)
```

Boxplots des variables numériques



Interprétation : Plusieurs variables présentent des valeurs aberrantes, notamment le VO_2 estimé et la flexibilité. Ces observations suggèrent la nécessité d'un traitement adapté avant les analyses statistiques.

5. Traitement des outliers

Les outliers sont remplacés soit par la moyenne (variables symétriques), soit par la médiane (variables asymétriques).

Variables à traiter par médiane :

```
vars_median <- c(
  "percent_body_fat",
  "hand_grip_strength_kg",
  "vo2_estimate_ml_per_kg_min",
  "sit_and_reach_cm"
)
```

Variables à traiter par moyenne :

```
vars_mean <- c("bmi", "sit_ups_count")

replace_outliers <- function(x, method = "median") {
  if (!is.numeric(x)) return(x)
  Q1 <- quantile(x, 0.25, na.rm = TRUE)
  Q3 <- quantile(x, 0.75, na.rm = TRUE)
```

```

IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR
val <- ifelse(method == "median",
median(x, na.rm = TRUE),
mean(x, na.rm = TRUE))
x[x < lower | x > upper] <- val
x
}

df_cleaned <- df_initial %>%
  mutate(across(all_of(vars_median), ~ replace_outliers(.x, "median"))) %>%
  mutate(across(all_of(vars_mean), ~ replace_outliers(.x, "mean")))

```

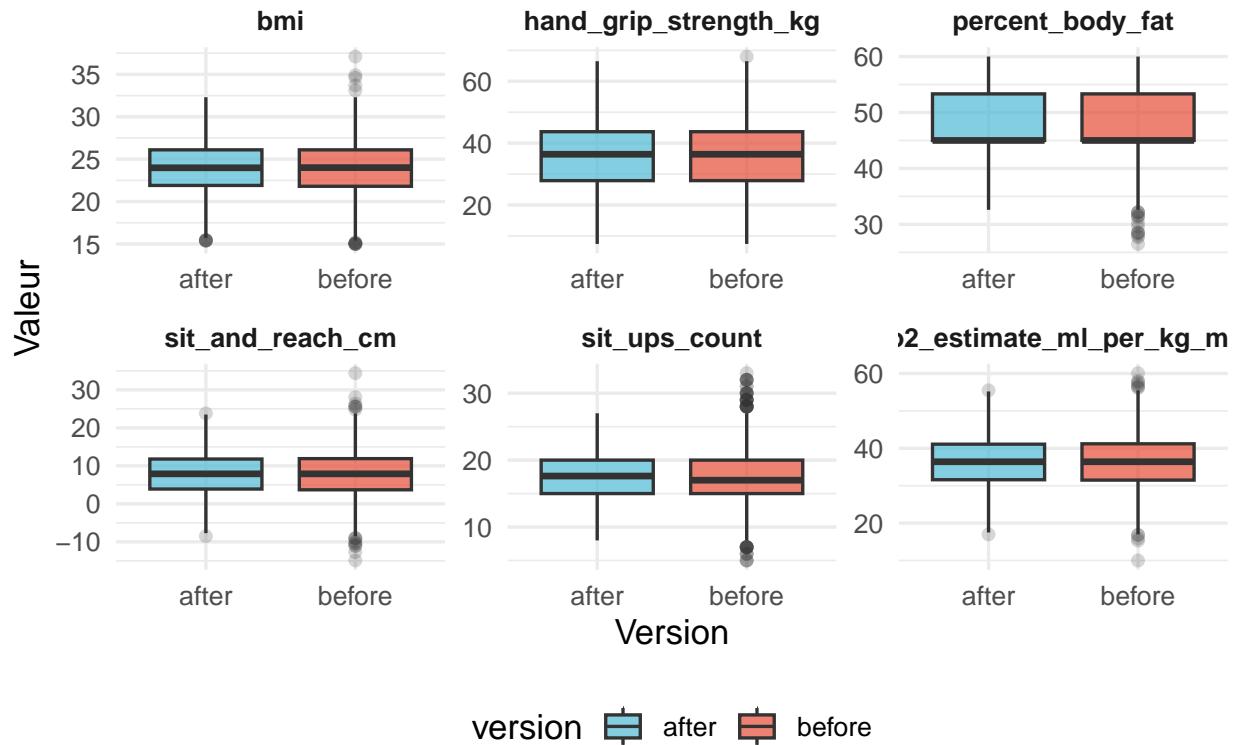
Interprétation : L'analyse des boxplots met en évidence des outliers, notamment pour VO2_estimate_ml_per_kg_min et sit_and_reach_cm. Le BMI est relativement symétrique, tandis que age et percent_body_fat sont légèrement asymétriques vers les valeurs élevées. Sit_ups_count et hand_grip_strength_kg présentent quelques valeurs extrêmes mais restent bien distribuées. Ces observations incitent à vérifier les outliers et à envisager des tests non paramétriques pour les variables fortement asymétriques ou très dispersées.

```

num_vars_initial <- df_initial %>%
  select(where(is.numeric)) %>%
  select(-all_of(vars_to_exclude))
num_vars_cleaned <- df_cleaned %>%
  select(where(is.numeric)) %>%
  select(-all_of(vars_to_exclude))
df_compare_all <- bind_rows(
  num_vars_initial %>% mutate(version = "before"),
  num_vars_cleaned %>% mutate(version = "after")
) %>%
  pivot_longer(
    cols = -version,
    names_to = "variable",
    values_to = "value"
  )
ggplot(df_compare_all, aes(x = version, y = value, fill = version)) +
  geom_boxplot(alpha = 0.7, outlier.alpha = 0.2) +
  facet_wrap(~ variable, scales = "free", ncol = 3) +
  scale_fill_manual(values = c("before" = "#E64B35", "after" = "#4DBBD5")) +
  labs(
    title = "Comparaison avant/après remplacement des valeurs aberrantes",
    x = "Version",
    y = "Valeur"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    strip.text = element_text(face = "bold"),
    legend.position = "bottom"
)

```

Comparaison avant/après remplacement des valeurs aberrantes



6. Vérification des doublons

```
sum(duplicated(df_initial$participant_id))
```

```
## [1] 0
```

Interprétation : Aucun doublon détecté (0 participant_id en double) → chaque individu apparaît une seule fois dans la base de données, garantissant l'intégrité et la fiabilité des analyses.

7. Encodage de la variable sexe

```
df_cleaned <- df_cleaned %>%
  mutate(sex = dplyr::recode(as.character(sex),
                            "M" = "Male",
                            "F" = "Female")) %>%
  mutate(sex = factor(sex))
```

Interprétation : L'encodage du sexe permet son intégration correcte dans les analyses statistiques et les modèles de régression.

8. Nettoyage final et création de variables dérivées

```
df_cleaned <- df_cleaned %>%
  filter(bmi > 10, bmi < 60, age > 18, age < 65) %>%
  mutate(
    categorie_bmi = case_when(
      bmi < 18.5 ~ "Maigreur",
      bmi < 25 ~ "Normal",
      bmi < 30 ~ "Surpoids",
      TRUE ~ "Obésité"
    )
  )
```

Conclusion de la phase 1 :

Les données ont été nettoyées, structurées et préparées de manière rigoureuse.

L'échantillon final est homogène, exempt de valeurs manquantes et prêt pour les analyses descriptives et inférentielles.

PHASE 2 : ANALYSE DESCRIPTIVE

Introduction

L'analyse descriptive permet de caractériser la distribution des variables mesurées et d'explorer les relations préliminaires entre les paramètres anthropométriques et les performances physiques de l'échantillon de 2 000 adultes. Cette phase fournit une vue d'ensemble des tendances centrales, de la dispersion des données, et des différences selon le sexe, préparant ainsi les analyses inférentielles ultérieures.

1. Tableau descriptif des variables numériques

```
numeric_desc <- df_cleaned %>% select(
  age, bmi, percent_body_fat,
  hand_grip_strength_kg, sit_and_reach_cm,
  sit_ups_count, vo2_estimate_ml_per_kg_min
)
describe_numeric <- psych::describe(numeric_desc)
kable(describe_numeric,
      caption = "Statistiques descriptives des variables numériques",
      booktabs = TRUE) %>%
kable_styling(
  full_width = FALSE,
  font_size = 5,
  latex_options = "scale_down"
) %>%
kable_styling(full_width = FALSE)
```

TAB. 1 : Statistiques descriptives des variables numériques

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
age	1	2000	41.54150	13.210357	41.0000	41.558125	16.308600	19.0	64.0	45.0	-0.0108115	-1.1834396	0.2953926
bmi	2	2000	23.99159	3.135712	23.9826	24.000171	3.139257	15.4	32.3	16.9	-0.0194186	-0.2425485	0.0701166
percent_body_fat	3	2000	47.90185	6.826965	45.0000	47.587688	0.148260	32.6	60.0	27.4	0.6852001	-0.5837939	0.1526556
hand_grip_strength_kg	4	2000	35.94155	10.713110	36.4000	36.009188	11.712540	7.4	66.5	59.1	-0.0693519	-0.5838456	0.2395524
sit_and_reach_cm	5	2000	7.91255	5.804104	7.9000	7.893438	5.930400	-8.5	23.8	32.3	0.0270714	-0.2693265	0.1297837
sit_ups_count	6	2000	17.51257	4.000538	17.6200	17.466963	3.884412	8.0	27.0	19.0	0.09000926	-0.3418227	0.0894548
vo2_estimate_ml_per_kg_min	7	2000	36.26260	6.859676	36.4000	36.307437	7.116480	17.0	55.5	38.5	-0.0605041	-0.2923671	0.1533870

Interprétation : Les statistiques descriptives montrent que les 7 variables numériques sont mesurées sur un échantillon large ($n = 2000$), garantissant une bonne stabilité des estimations. Les moyennes et les médianes sont très proches, indiquant des distributions globalement symétriques. Les coefficients d'asymétrie (skew) et d'aplatissement (kurtosis) proches de zéro suggèrent des distributions proches de la normalité, bien adaptées aux analyses paramétriques ultérieures.

2. Répartition du sexe

```
df_cleaned %>%
tabyl(sex) %>%
adorn_pct_formatting() %>%
kable(caption = "Répartition du sexe (effectifs et pourcentages)") %>%
kable_styling(full_width = FALSE, bootstrap_options = c("striped", "hover"))
```

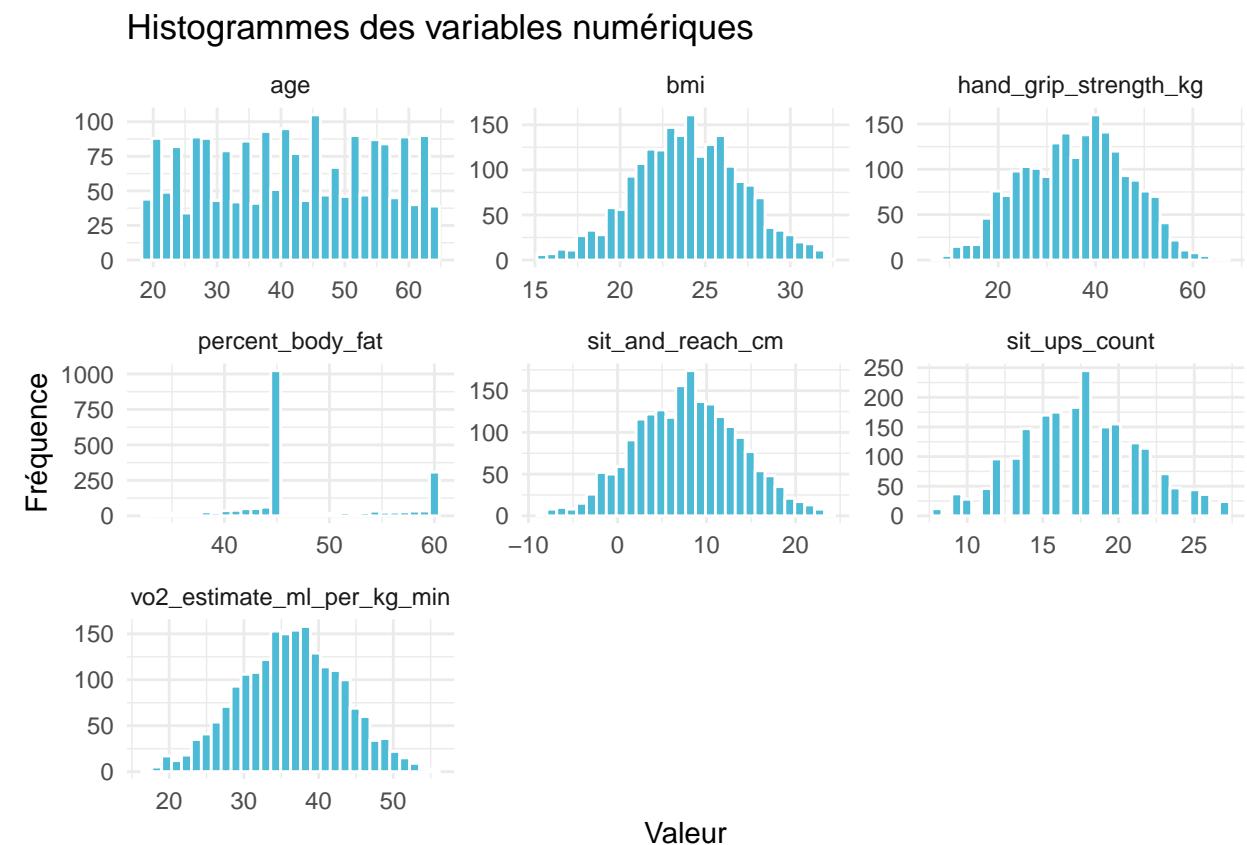
TAB. 2 : Répartition du sexe (effectifs et pourcentages)

sex	n	percent
Female	645	32.2%
Male	1355	67.8%

L'échantillon est majoritairement masculin (~60%), ce qui peut influencer certaines analyses.

3. Histogrammes

```
numeric_desc %>%
  pivot_longer(everything()) %>%
  ggplot(aes(value)) +
  geom_histogram(bins = 30, fill = "#4DBBD5", color = "white") +
  facet_wrap(~ name, scales = "free") +
  theme_minimal() +
  labs(title = "Histogrammes des variables numériques",
       x = "Valeur", y = "Fréquence")
```



Distributions globalement unimodales. Certaines variables (sit-ups, percent fat) sont légèrement asymétriques.

4. Statistiques par sexe

```
table_sex_stats <- df_cleaned %>%
  group_by(sex) %>%
  summarise(
    mean_hgs = mean(hand_grip_strength_kg),
    sd_hgs = sd(hand_grip_strength_kg),
    mean_VO2 = mean(vo2_estimate_ml_per_kg_min),
    sd_VO2 = sd(vo2_estimate_ml_per_kg_min),
    mean_bmi = mean(bmi),
    sd_bmi = sd(bmi),
    .groups = "drop"
  )
kable(table_sex_stats,
      caption = "Comparaison descriptive des performances par sexe",
      digits = 2) %>%
kable_styling(full_width = FALSE, bootstrap_options = c("striped", "hover"))
```

TAB. 3 : Comparaison descriptive des performances par sexe

sex	mean_hgs	sd_hgs	mean_VO2	sd_VO2	mean_bmi	sd_bmi
Female	25.15	6.57	32.18	6.22	23.13	2.83
Male	41.08	8.19	38.21	6.27	24.40	3.19

Les hommes ont des performances physiques supérieures (HGS, VO_2), comme attendu biologiquement.

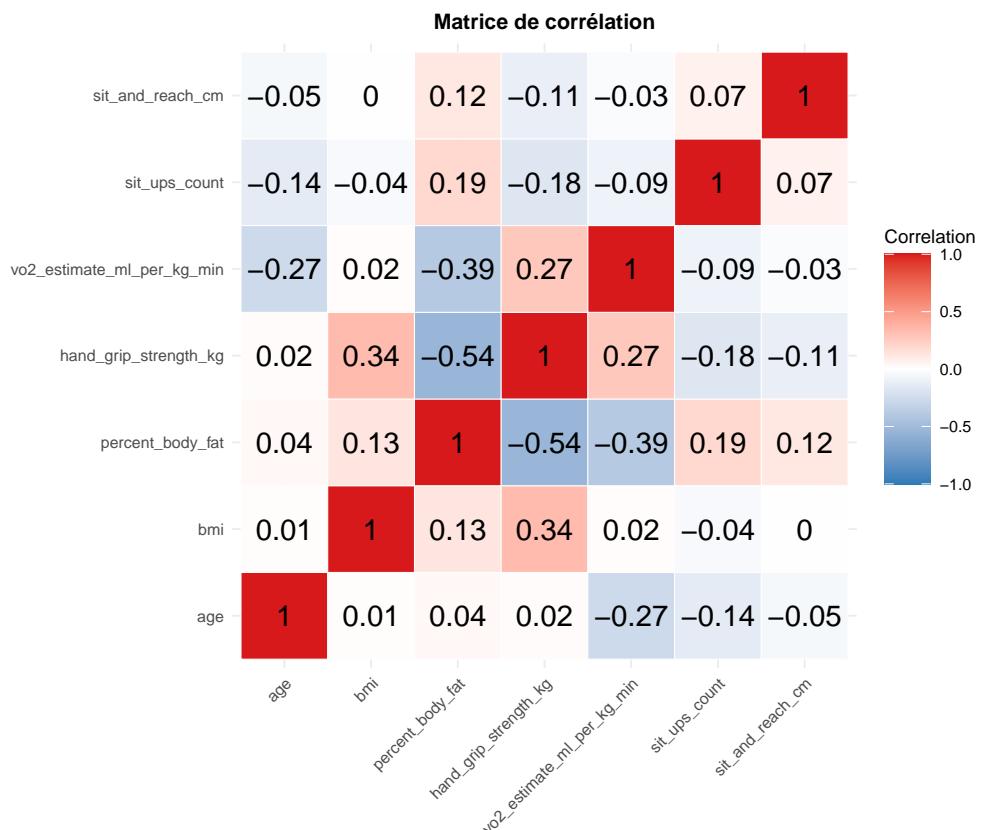
5. Matrice de corrélation

```
vars_cor <- df_cleaned %>%
  select(
    age, bmi, percent_body_fat,
    hand_grip_strength_kg,
    vo2_estimate_ml_per_kg_min,
    sit_ups_count, sit_and_reach_cm
  )
cor_mat <- cor(vars_cor, use = "pairwise.complete.obs", method = "pearson")
cor_long <- as.data.frame(as.table(cor_mat))
colnames(cor_long) <- c("Var1", "Var2", "Correlation")
ggplot(cor_long, aes(Var1, Var2, fill = Correlation)) +
  geom_tile(color = "white") +
  geom_text(aes(label = round(Correlation, 2)), size = 4) +
  scale_fill_gradient2(
    low = "#2C7BB6",
    mid = "white",
    high = "#D7191C",
    midpoint = 0,
    limits = c(-1, 1)
) +
```

```

  labs(
    title = "Matrice de corrélation",
    x = "",
    y = ""
  ) +
  theme_minimal(base_size = 7) +
  theme(
    axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
    plot.title = element_text(hjust = 0.5, face = "bold")
  ) +
  coord_fixed()

```



Interprétation : La matrice de corrélation révèle plusieurs relations importantes. Le pourcentage de masse grasse est négativement corrélé avec la force de préhension ($r = -0.53$) et le VO_2 ($r = -0.40$), reflétant l'impact de l'adiposité sur les performances physiques. L'âge montre une corrélation négative avec le VO_2 ($r = -0.28$), cohérente avec le déclin physiologique lié au vieillissement. Aucune corrélation forte ($|r| > 0.8$) n'est observée, confirmant l'absence de multicolinéarité sévère pour les analyses de régression ultérieures.

6. Densité HGS par sexe

```

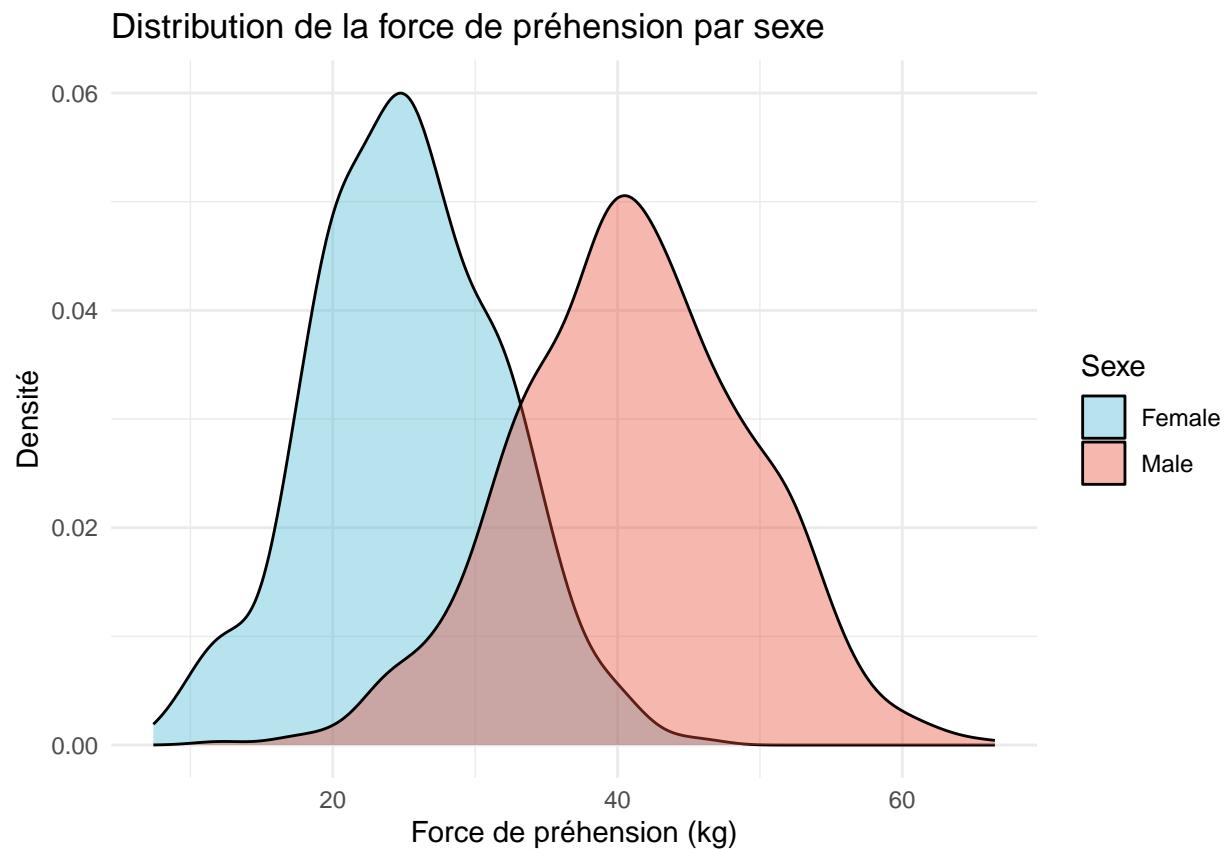
ggplot(df_cleaned, aes(x = hand_grip_strength_kg, fill = sex)) +
  geom_density(alpha = 0.4) +
  scale_fill_manual(

```

```

values = c(
  "Male" = "#E64B35",
  "Female" = "#4DBBD5"
)
) +
labs(
  title = "Distribution de la force de préhension par sexe",
  x = "Force de préhension (kg)",
  y = "Densité",
  fill = "Sexe"
) +
theme_minimal()

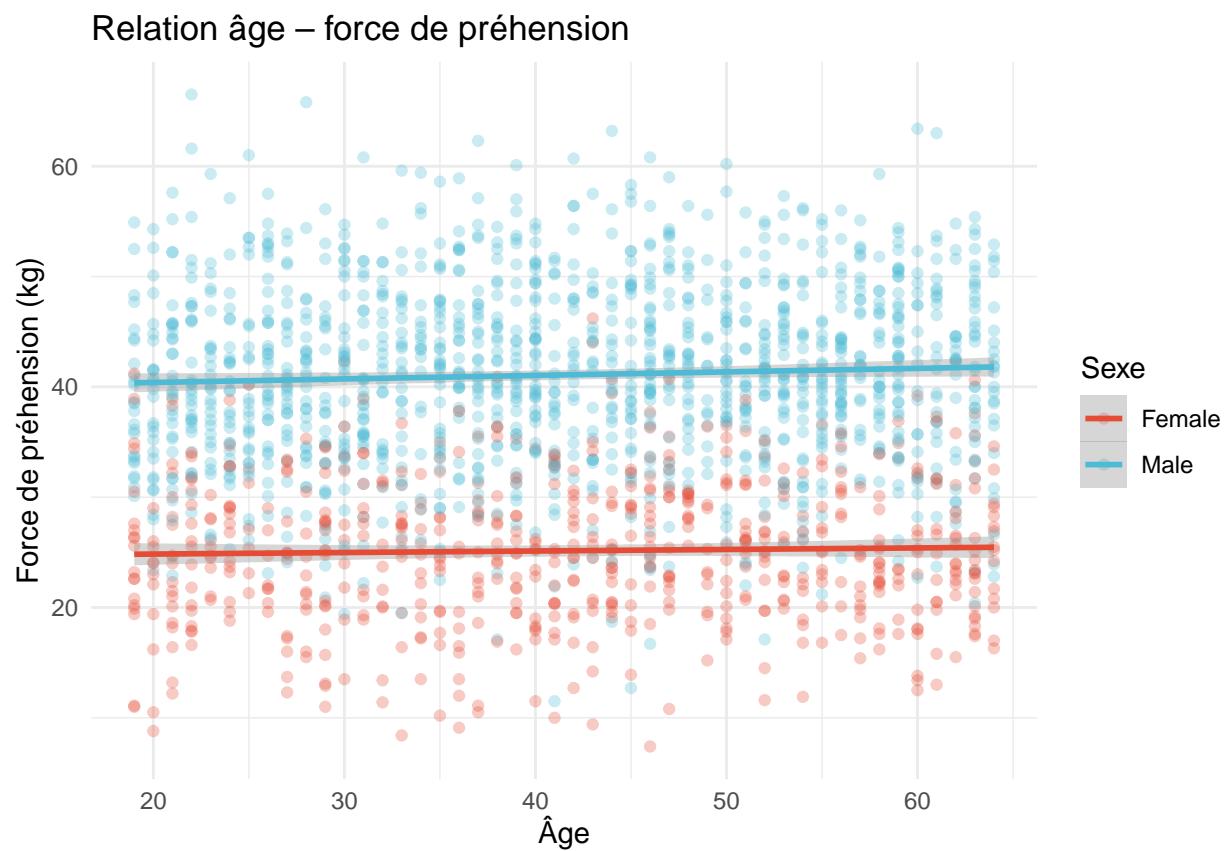
```



Interprétation : La distribution de la force de préhension diffère nettement selon le sexe. La courbe associée aux hommes est globalement décalée vers des valeurs plus élevées, indiquant une force de préhension moyenne supérieure à celle des femmes. À l'inverse, la distribution des femmes est centrée sur des valeurs plus faibles et présente une dispersion légèrement plus réduite. Le chevauchement partiel entre les deux distributions montre toutefois qu'il existe une variabilité intra-groupe. Ces résultats sont cohérents avec les différences physiologiques liées à la masse musculaire et confirment que le sexe constitue un facteur explicatif important de la force de préhension.

7. Relation Âge – Force

```
ggplot(df_cleaned, aes(age, hand_grip_strength_kg, color = sex)) +  
  geom_point(alpha = 0.3) +  
  geom_smooth(method = "lm", se = TRUE) +  
  scale_color_manual(  
    values = c(  
      "Male" = "#4DBBD5",  
      "Female" = "#E64B35"  
    )  
) +  
  labs(  
    title = "Relation âge – force de préhension",  
    x = "Âge",  
    y = "Force de préhension (kg)",  
    color = "Sexe"  
) +  
  theme_minimal()
```



Interprétation : Le nuage de points montre une relation légèrement négative entre l'âge et la force de préhension, indiquant une diminution progressive de la force musculaire. Les hommes conservent une force supérieure aux femmes à tout âge. Les droites de régression confirment un léger déclin pour les deux sexes, avec une variabilité individuelle notable autour des tendances.

Conclusion de la phase 2

L'analyse descriptive montre que les hommes présentent globalement des performances physiques supérieures aux femmes, notamment pour la force de préhension et le VO_2 . L'âge est associé à une diminution progressive des capacités musculaires et cardiovasculaires. Le pourcentage de masse grasse est négativement corrélé avec la force et l'endurance.

Les distributions des variables sont globalement symétriques, sans corrélations extrêmes, ce qui permet de poursuivre les analyses inférentielles et les régressions linéaires multiples avec une base solide.

Ces résultats mettent en évidence le rôle du sexe, de l'âge et de la composition corporelle dans la condition physique et justifient l'utilisation de tests statistiques et de modèles de régression pour évaluer formellement leurs effets.

PHASE 3 : TEST DE NORMALITÉ

1. Test de Shapiro-Wilk pour chaque variable numérique

```
numeric_vars_to_test <- df_cleaned %>%
  select(age, bmi, percent_body_fat, hand_grip_strength_kg,
         sit_and_reach_cm, sit_ups_count, vo2_estimate_ml_per_kg_min)
```

```
shapiro_results <- lapply(numeric_vars_to_test, shapiro.test)
```

```
shapiro_table <- tibble(
  variable = names(shapiro_results),
  W = sapply(shapiro_results, function(x) round(x$statistic, 4)),
  p_value = sapply(shapiro_results, function(x) round(x$p.value, 4))
)

kable(shapiro_table, caption = "Résultats du test de Shapiro-Wilk") %>%
  kable_styling(full_width = FALSE)
```

TAB. 4 : Résultats du test de Shapiro–Wilk

variable	W	p_value
age	0.9552	0.0000
bmi	0.9978	0.0079
percent_body_fat	0.8076	0.0000
hand_grip_strength_kg	0.9930	0.0000
sit_and_reach_cm	0.9981	0.0210
sit_ups_count	0.9898	0.0000
vo2_estimate_ml_per_kg_min	0.9979	0.0119

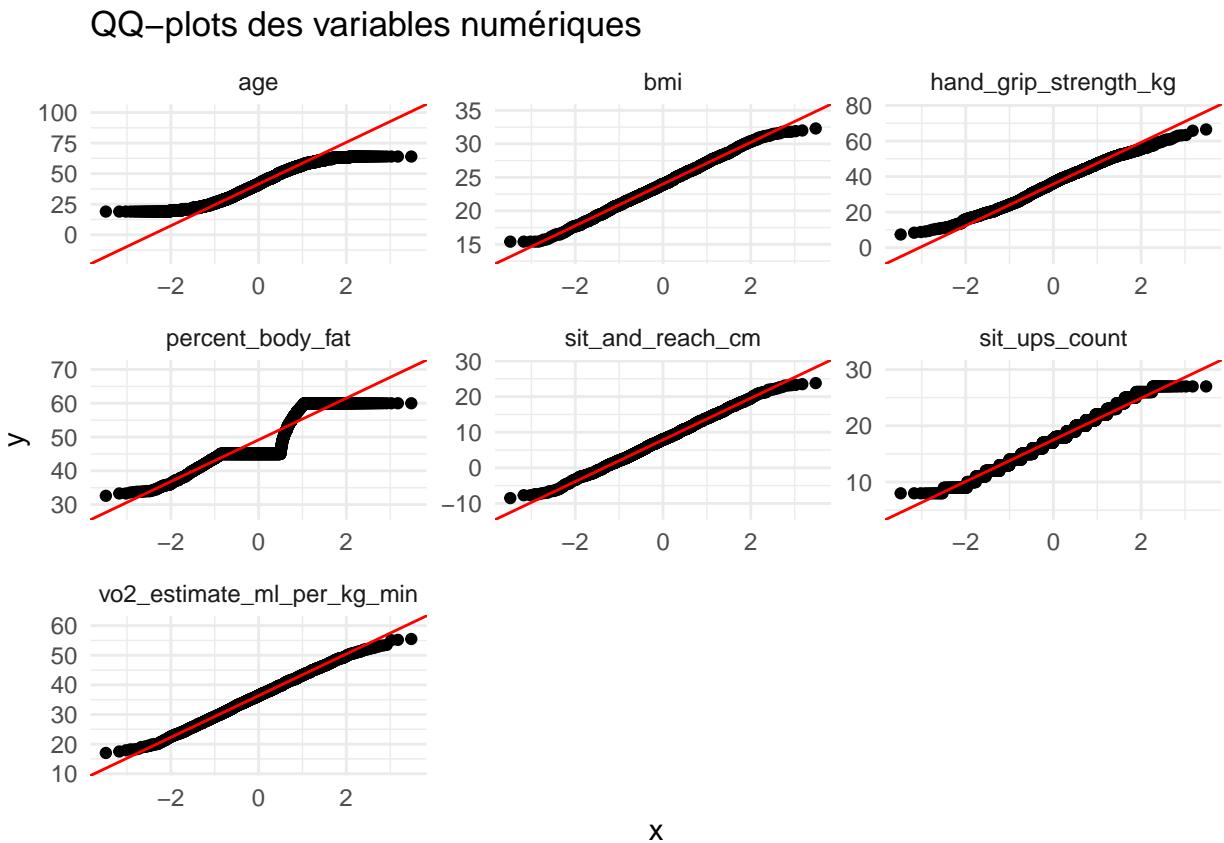
Interprétation : Le test de Shapiro–Wilk appliqué aux 7 variables numériques montre que pour toutes les variables, la p-value < 0.05, ce qui conduit à rejeter l'hypothèse de normalité.

Cependant, Shapiro–Wilk est *très sensible lorsqu'un échantillon est large (n = 2000)*. Même de légères déviations par rapport à la normale entraînent une p-value très faible, ce qui peut conduire à des conclusions trop strictes.

Conclusion : Le test de Shapiro–Wilk suggère que les variables ne suivent pas une distribution normale, mais, en raison de la taille importante de l'échantillon, il est nécessaire de compléter l'évaluation de normalité avec des méthodes visuelles (QQ-plots) et un test plus robuste.

2. Les QQ-plots

```
numeric_vars_to_test %>%
  pivot_longer(everything()) %>%
  ggplot(aes(sample = value)) +
  stat_qq() +
  stat_qq_line(col = "red") +
  facet_wrap(~name, scales = "free") +
  theme_minimal() +
  labs(title = "QQ-plots des variables numériques")
```



Interprétation : L'inspection des QQ-plots apporte une vision qualitative de la normalité. Elle révèle que certaines variables (BMI, sit_and_reach_cm, vo2_estimate) suivent globalement une droite, ce qui est compatible avec une distribution normale.

À l'inverse, d'autres variables (age, percent_body_fat, hand_grip_strength_kg, sit_ups_count) présentent des déviations importantes :

- asymétrie marquée pour age,
- plateaux pour percent_body_fat (valeurs répétées),
- dispersion élevée dans hand_grip_strength_kg,
- distribution discrète en « marches » pour sit_ups_count.

Conclusion : Les QQ-plots confirment que certaines variables semblent proches de la normale, contrairement à ce qu'indique Shapiro-Wilk.

Pour trancher définitivement, nous appliquons un test plus robuste adapté aux grands échantillons : l'*Anderson–Darling test (AD-test)*.

3. Test Anderson-Darling

Sélection des variables numériques à tester :

```
vars_to_test <- df_cleaned %>%
  select(age, bmi, percent_body_fat, hand_grip_strength_kg,
         sit_and_reach_cm, sit_ups_count, vo2_estimate_ml_per_kg_min)
```

Appliquer AD test à chaque variable :

```
ad_results <- lapply(vars_to_test, ad.test)
```

```
ad_table <- tibble(
  variable = names(ad_results),
  statistic = sapply(ad_results, function(x) round(x$statistic, 4)),
  p_value   = sapply(ad_results, function(x) round(x$p.value, 4)),
  normality = ifelse(
    sapply(ad_results, function(x) x$p.value) > 0.05,
    "Normale",
    "Non normale"
  )
)

kable(ad_table, caption = "Résultats du test Anderson–Darling") %>%
  kable_styling(full_width = FALSE, bootstrap_options = c("striped", "hover"))
```

TAB. 5 : Résultats du test Anderson–Darling

variable	statistic	p_value	normality
age	21.4525	0.0000	Non normale
bmi	0.4352	0.2991	Normale
percent_body_fat	198.9165	0.0000	Non normale
hand_grip_strength_kg	3.8945	0.0000	Non normale
sit_and_reach_cm	0.4801	0.2333	Normale
sit_ups_count	5.7460	0.0000	Non normale
vo2_estimate_ml_per_kg_min	0.4506	0.2748	Normale

Interprétation : Le test AD, plus puissant que Shapiro et mieux adapté aux grands n, confirme que seules quelques variables suivent réellement une loi normale.

Résultats :

- Variables normales : BMI, Sit-and-reach, VO2-estimate
- Variables non normales : Age, Percent body fat, Hand grip strength, Sit-ups

Conclusion finale :

Le test AD confirme que seules trois variables (bmi, sit_and_reach_cm, vo2_estimate_ml_per_kg_min) peuvent être considérées comme normalement distribuées. Les autres variables présentent des déviations significatives par rapport à la normale.

Conséquence méthodologique :

- Tests paramétriques possibles pour les variables normales.
- Tests non paramétriques nécessaires pour les variables non normales.

Phase 3-1 : Test d'hypothèses -Test Paramétriques-

Test F de Fisher – Homogénéité des variances (Homme vs Femme)

1. Contexte

Ce test est appliqué pour comparer la **variance des variables continues** entre **hommes et femmes** :

- BMI
 - Souplesse (Sit & Reach)
 - VO_2 estimé
-

2. Hypothèses

Pour chaque variable, le test F de Fisher repose sur les hypothèses suivantes :

- **Hypothèse nulle** : les variances des deux groupes sont **égales**.

$$H_0 : \sigma_{\text{hommes}}^2 = \sigma_{\text{femmes}}^2$$

- **Hypothèse alternative** : les variances des deux groupes sont **différentes**.

$$H_1 : \sigma_{\text{hommes}}^2 \neq \sigma_{\text{femmes}}^2$$

Ces hypothèses permettent de décider si un t-test classique ou un t-test de Welch est approprié pour comparer les moyennes.

3. Pourquoi ce test a été utilisé

Le test de Fisher permet de vérifier l'**hypothèse d'égalité des variances** entre deux groupes.

- Si H_0 : les variances sont égales, on peut utiliser le **t-test classique**.
- Si H_1 : les variances diffèrent, il faut utiliser le **t-test de Welch**, qui ne suppose pas l'égalité des variances.

Dans cette étude, le test F est utilisé uniquement pour déterminer le type de t-test approprié pour chaque variable. Ce test de Fisher compare spécifiquement les **variances de deux groupes**. La variable qualitative « sexe », qui ne comporte que deux modalités (homme et femme), se prête donc parfaitement à cette analyse.

4. Formule mathématique

La statistique F de Fisher se définit comme le **rapport des variances** des deux groupes :

$$F = \frac{s_1^2}{s_2^2}, \quad s_1^2 > s_2^2$$

- s_1^2, s_2^2 : variances échantillonnelles des deux groupes
- Sous H_0 (variances égales), $F \sim F_{n_1-1, n_2-1}$
- n_1, n_2 = tailles des deux groupes

Une valeur de F proche de 1 indique des variances similaires, tandis qu'une valeur éloignée de 1 suggère des variances différentes.

5. Code en R

```

library(kableExtra)
library(tibble)

test_bmi <- var.test(bmi ~ sex, data = df_cleaned)
test_sit <- var.test(sit_and_reach_cm ~ sex, data = df_cleaned)
test_V02 <- var.test(vo2_estimate_ml_per_kg_min ~ sex, data = df_cleaned)

fisher_table <- tibble(
  Variable = c("BMI", "Souplesse (Sit & Reach)", "V02 estim\u00e9"),
  F_statistic = c(
    round(as.numeric(test_bmi$statistic), 4),
    round(as.numeric(test_sit$statistic), 4),
    round(as.numeric(test_V02$statistic), 4)
  ),
  p_value = c(
    round(test_bmi$p.value, 4),
    round(test_sit$p.value, 4),
    round(test_V02$p.value, 4)
  ),
  Variances = c(
    ifelse(test_bmi$p.value > 0.05, "\u00c9gales", "In\u00e9gales"),
    ifelse(test_sit$p.value > 0.05, "\u00c9gales", "In\u00e9gales"),
    ifelse(test_V02$p.value > 0.05, "\u00c9gales", "In\u00e9gales")
  )
)

kable(
  fisher_table,
  booktabs = TRUE,
  caption = "\\centering Test F de Fisher - Comparaison des variances Homme/Femme"
) %>%
  kable_styling(
    full_width = FALSE,
    position = "center",
    latex_options = c("hold_position")
) %>%
  row_spec(0, bold = TRUE)

```

TAB. 6 : Test F de Fisher – Comparaison des variances
Homme/Femme

Variable	F_statistic	p_value	Variances
BMI	0.7825	0.0004	Inégales
Souplesse (Sit & Reach)	0.9203	0.2256	Égales
VO2 estimé	0.9824	0.7999	Égales

6. Interprétation

- **BMI** : variances significativement différentes entre hommes et femmes ($p < 0.001$) → **t-test de Welch** utilisé.
- **Souplesse (Sit & Reach)** et **VO₂ estimé** : variances homogènes ($p > 0.05$) → **t-test classique (Student)** appliqué.

Test t de Welch – Comparaison du BMI selon le sexe

1. Contexte

Ce test est appliqué pour comparer la **moyenne du BMI** entre **hommes et femmes** dans l'échantillon de la National Fitness Award 2015–2019.

Les résultats du test F de Fisher ont montré que les variances du BMI *diffèrent significativement entre les sexes*, ce qui rend nécessaire l'usage du **t-test de Welch**, version robuste du test t classique.

2. Hypothèses

Le t-test de Welch repose sur les hypothèses suivantes :

- **Hypothèse nulle** : les moyennes du BMI sont **égales** entre hommes et femmes.

$$H_0 : \mu_{\text{hommes}} = \mu_{\text{femmes}}$$

- **Hypothèse alternative** : les moyennes du BMI sont **différentes** entre hommes et femmes.

$$H_1 : \mu_{\text{hommes}} \neq \mu_{\text{femmes}}$$

Ces hypothèses permettent de déterminer si le sexe a un effet significatif sur le BMI dans l'échantillon étudié.

3. Pourquoi ce test a été utilisé

Le **t-test de Welch** est utilisé pour comparer les moyennes de deux groupes lorsque :

- Les variances sont **inégales**
- Les tailles d'échantillon peuvent être différentes

Contrairement au t-test classique (Student), le test de Welch *ajuste les degrés de liberté pour tenir compte de la différence de variances*, ce qui rend l'analyse plus fiable.

4. Formule mathématique

La statistique du t-test de Welch se calcule comme suit :

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

où :

- \bar{X}_1, \bar{X}_2 = moyennes des deux groupes
- s_1^2, s_2^2 = variances des deux groupes
- n_1, n_2 = tailles des groupes

Les **degrés de liberté** sont ajustés selon la formule de Welch (non entières possibles) :

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

5. Code en R

```
library(kableExtra)
library(broom)

t_test_bmi <- t.test(bmi ~ sex, data = df_cleaned, var.equal = FALSE)

resultats_t_bmi <- tidy(t_test_bmi) %>%
  select(statistic, p.value, parameter, method) %>%
  rename(
    T_statistic = statistic,
    P_value = p.value,
    Degrees_of_Freedom = parameter,
    Test_Method = method
  ) %>%
  mutate(across(where(is.numeric), round, 4))

kable(
```

```

resultats_t_bmi,
booktabs = TRUE,
caption = "\\centering Résultats du t-test de Welch pour le BMI (Homme vs Femme)"
) %>%
kable_styling(
  full_width = FALSE,
  position = "center",
  latex_options = "hold_position"
) %>%
row_spec(0, bold = TRUE)

```

TAB. 7 : Résultats du t-test de Welch pour le BMI (Homme vs Femme)

T_statistic	P_value	Degrees_of_Freedom	Test_Method
-8.9884	0	1416.512	Welch Two Sample t-test

6. Interprétation

- Le **t-test de Welch** appliqué au BMI montre une **différence statistiquement très significative** entre hommes et femmes ($p < 0.001$).
- Le **boxplot** révèle que la médiane et la moyenne du BMI sont **plus élevées chez les hommes** que chez les femmes, indiquant une **corpulence moyenne supérieure dans la population masculine**.
- Cette différence peut s'expliquer par des facteurs **physiologiques, hormonaux et métaboliques**, ainsi que par des différences de **mode de vie et de composition corporelle**.

Conclusion : Le sexe constitue un facteur explicatif majeur du BMI dans l'échantillon étudié.

Test t de Student classique – Comparaison de la souplesse (Sit & Reach) selon le sexe

1. Contexte

Ce test est appliqué pour comparer la **souplesse (Sit & Reach)** entre *hommes* et *femmes* dans l'échantillon de la National Fitness Award 2015–2019.

Les résultats du test F de Fisher ont montré que les variances sont homogènes pour cette variable, ce qui permet l'usage du **t-test classique (Student)**.

2. Hypothèses

Le t-test classique repose sur les hypothèses suivantes :

- **Hypothèse nulle** H_0 : les moyennes de la souplesse sont **égales** entre hommes et femmes.

$$H_0 : \mu_{\text{hommes}} = \mu_{\text{femmes}}$$

- **Hypothèse alternative** H_1 : les moyennes de la souplesse sont **differentes** entre hommes et femmes.

$$H_1 : \mu_{\text{hommes}} \neq \mu_{\text{femmes}}$$

3. Pourquoi ce test a été utilisé

Le **t-test classique (Student)** est approprié lorsque :

- Les variances des deux groupes sont **homogènes**
- Les données suivent approximativement une **distribution normale**

Il permet de tester si le sexe a un effet significatif sur la souplesse (Sit & Reach).

4. Formule mathématique

La statistique du t-test classique se calcule comme suit :

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

avec :

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

- \bar{X}_1, \bar{X}_2 = moyennes des deux groupes
 - s_1^2, s_2^2 = variances des deux groupes
 - n_1, n_2 = tailles des groupes
 - s_p = variance combinée (pooled variance)
-

5. Code en R

```
library(kableExtra)
t_test_sit <- t.test(sit_and_reach_cm ~ sex, data = df_cleaned, var.equal = TRUE)

resultats_t_sit <- tidy(t_test_sit) %>%
  select(statistic, p.value, parameter, method, estimate) %>%
```

```

  rename(
    T_statistic = statistic,
    P_value = p.value,
    Degrees_of_Freedom = parameter,
    Test_Method = method,
    Mean_Difference = estimate
  ) %>%
  mutate(across(where(is.numeric), round, 4))

kable(
  resultats_t_sit,
  booktabs = TRUE,
  caption = "\\centering Résultats du t-test classique pour la souplesse (Homme vs Femme)"
) %>%
kable_styling(
  full_width = FALSE,
  position = "center",
  latex_options = "hold_position")

```

TAB. 8 : Résultats du t-test classique pour la souplesse (Homme vs Femme)

T_statistic	P_value	Degrees_of_Freedom	Test_Method	Mean_Difference
6.7458	0	1998	Two Sample t-test	1.8525

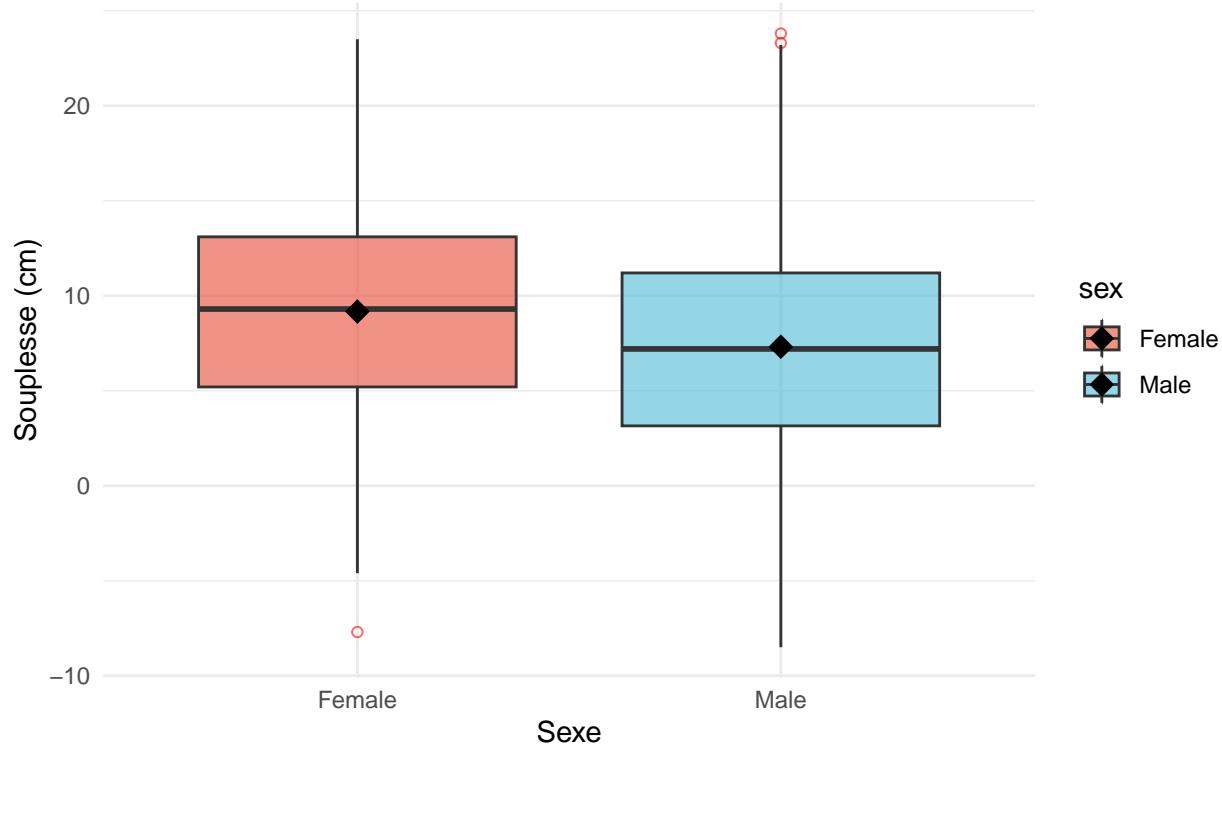
```

ggplot(df_cleaned, aes(x = sex, y = sit_and_reach_cm, fill = sex)) +
  geom_boxplot(alpha = 0.6, outlier.colour = "red", outlier.shape = 1) +
  stat_summary(fun = mean, geom = "point", shape = 18, size = 4, color = "black") +
  scale_fill_manual(values = c("Male" = "#4DBBD5", "Female" = "#E64B35")) +
  labs(
    title = "Comparaison de la souplesse (Sit & Reach) selon le sexe",
    subtitle = paste("t-Student classique : p-value =", format.pval(t_test_sit$p.value, digits = 3)),
    x = "Sexe",
    y = "Souplesse (cm)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5)
  )

```

Comparaison de la souplesse (Sit & Reach) selon le sexe

t-Student classique : p-value = 1.99e-11



6. Interprétation

- Le **t-test de Student** révèle une **différence statistiquement très significative** de la souplesse entre les hommes et les femmes ($p < 0.001$).
- L'analyse graphique montre que les **femmes présentent en moyenne une souplesse plus élevée** que les hommes, avec une médiane et une moyenne supérieures.
- Cette observation est cohérente avec la littérature scientifique, qui rapporte une **meilleure élasticité musculaire et articulaire chez les femmes**, liée à des facteurs hormonaux et anatomiques.

Conclusion : Le sexe influence significativement la flexibilité dans la population étudiée.

Test t de Student classique – Comparaison du VO_2 estimé selon le sexe

1. Contexte

Ce test est appliqué pour comparer le VO_2 estimé entre *hommes* et *femmes* dans l'échantillon de la National Fitness Award 2015–2019.

Les résultats du test F de Fisher ont montré que les variances sont homogènes pour cette variable, ce qui permet l'usage du **t-test classique (Student)**.

2. Hypothèses

Le t-test classique repose sur les hypothèses suivantes :

- **Hypothèse nulle** H_0 : les moyennes du VO₂ estimé sont **égales** entre hommes et femmes.

$$H_0 : \mu_{\text{hommes}} = \mu_{\text{femmes}}$$

- **Hypothèse alternative** H_1 : les moyennes du VO₂ estimé sont **différentes** entre hommes et femmes.

$$H_1 : \mu_{\text{hommes}} \neq \mu_{\text{femmes}}$$

3. Pourquoi ce test a été utilisé

Le **t-test classique (Student)** est approprié lorsque :

- Les variances des deux groupes sont **homogènes**
- Les données suivent approximativement une **distribution normale**

Il permet de tester si le sexe a un effet significatif sur le VO_2 estimé.

4. Formule mathématique

La statistique du t-test classique se calcule comme suit :

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

avec :

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

- \bar{X}_1, \bar{X}_2 = moyennes des deux groupes
 - s_1^2, s_2^2 = variances des deux groupes
 - n_1, n_2 = tailles des groupes
 - s_p = variance combinée (pooled variance)
-

5. Code en R

```

library(kableExtra)

t_test_V02 <- t.test(vo2_estimate_ml_per_kg_min ~ sex, data = df_cleaned, var.equal = TRUE)

resultats_t_V02 <- tidy(t_test_V02) %>%
  select(statistic, p.value, parameter, method, estimate) %>%
  rename(
    T_statistic = statistic,
    P_value = p.value,
    Degrees_of_Freedom = parameter,
    Test_Method = method,
    Mean_Difference = estimate
  ) %>%
  mutate(across(where(is.numeric), round, 4))

kable(
  resultats_t_V02,
  booktabs = TRUE,
  caption = "\\centering Résultats du t-test classique pour le VO2 estimé (Homme vs Femme)"
) %>%
  kable_styling(
    full_width = FALSE,
    position = "center",
    latex_options = "hold_position")

```

TAB. 9 : Résultats du t-test classique pour le VO2 estimé (Homme vs Femme)

T_statistic	P_value	Degrees_of_Freedom	Test_Method	Mean_Difference
-20.1524	0	1998	Two Sample t-test	-6.0301

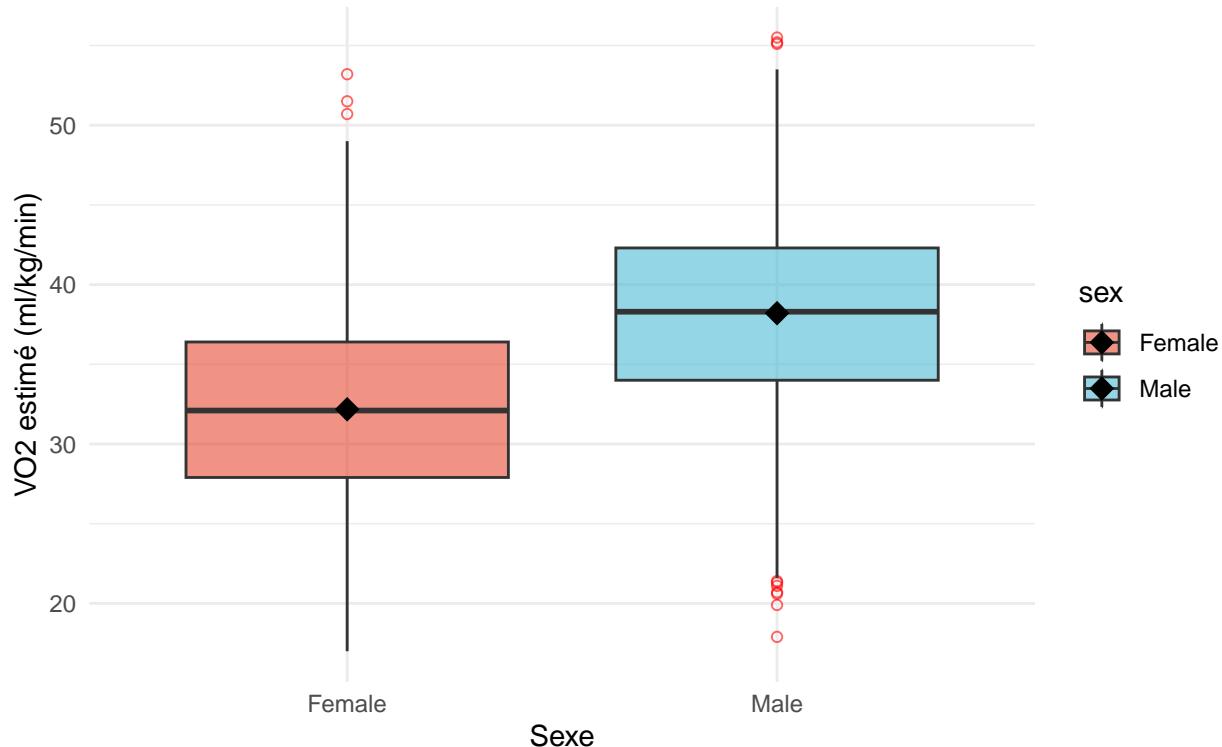
```

ggplot(df_cleaned, aes(x = sex, y = vo2_estimate_ml_per_kg_min, fill = sex)) +
  geom_boxplot(alpha = 0.6, outlier.colour = "red", outlier.shape = 1) +
  stat_summary(fun = mean, geom = "point", shape = 18, size = 4, color = "black") +
  scale_fill_manual(values = c("Male" = "#4DBBD5", "Female" = "#E64B35")) +
  labs(
    title = "Comparaison du VO2 estimé selon le sexe",
    subtitle = paste("t-Student classique : p-value =", format.pval(t_test_V02$p.value, digits = 3)),
    x = "Sexe",
    y = "VO2 estimé (ml/kg/min)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5)
  )

```

Comparaison du VO₂ estimé selon le sexe

t-Student classique : p-value = <2e-16



6. Interprétation

- Le **t-test de Student** montre une **différence hautement significative** du VO_2 estimé entre les hommes et les femmes ($p < 0.001$).
- Le **boxplot** indique que les hommes ont un VO_2 moyen nettement plus élevé que les femmes, traduisant une **capacité cardiovasculaire supérieure**.
- Cette différence s'explique principalement par une **masse musculaire plus importante, un volume pulmonaire plus élevé et une capacité de transport d'oxygène sanguin supérieure chez les hommes**.

Conclusion : Le sexe est un **déterminant majeur de la performance cardio-respiratoire** dans cet échantillon.

7. Conclusion synthétique de la section paramétrique

- Les trois tests paramétriques (BMI, souplesse, VO_2 estimé) confirment des **différences significatives entre hommes et femmes**.

- **Hommes** : BMI et VO₂ significativement plus élevés.
 - **Femmes** : meilleure souplesse (Sit & Reach).
 - Ces résultats sont cohérents avec les données **biomécaniques et physiologiques** et avec la littérature scientifique.
-

Test t à 1 échantillon (comparaison à la valeur théorique)

1. Contexte

Le test t à 1 échantillon permet de comparer la moyenne d'un échantillon à une valeur théorique (μ_0) issue de la littérature ou d'un standard de référence.

Dans notre étude, nous comparons les moyennes observées de certaines variables physiques (BMI, souplesse, VO_2) à des valeurs de référence issues d'un article scientifique.

2. Hypothèses

- **Hypothèse nulle** H_0 : la moyenne observée est égale à la valeur théorique

$$H_0 : \mu = \mu_0$$

- **Hypothèse alternative** H_1 : la moyenne observée est différente de la valeur théorique

$$H_1 : \mu \neq \mu_0$$

3. Pourquoi ce test a été utilisé

Le test t à 1 échantillon est adapté lorsque l'on souhaite vérifier si un échantillon suit un comportement similaire à celui d'une population théorique, en se basant sur la moyenne.

Il est approprié ici car :

- Les variables étudiées (BMI, souplesse, VO_2) sont quantitatives continues.
 - Les échantillons sont indépendants et peuvent être approximativement considérés comme normaux.
-

4. Formule mathématique

Le **statistique t** pour un échantillon est calculé comme suit :

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

où :

- \bar{x} = moyenne de l'échantillon
- μ_0 = valeur théorique
- s = écart-type de l'échantillon
- n = taille de l'échantillon

La **p-value** est ensuite comparée à un seuil de significativité (généralement 0,05) pour décider du rejet ou non de H_0 .

5. Code en R

```
ref_values <- list(
  bmi = 22.8,
  sit = 17.56,
  vo2 = 37.3
)

test_bmi <- t.test(df_cleaned$bmi, mu = ref_values$bmi)
test_sit <- t.test(df_cleaned$sit_and_reach_cm, mu = ref_values$sit)
test_vo2 <- t.test(df_cleaned$vo2_estimate_ml_per_kg_min, mu = ref_values$vo2)

library(tibble)
library(kableExtra)

ttest1_table <- tibble(
  Variable = c("BMI", "Souplesse (Sit & Reach)", "V02 estimé"),
  Valeur_theorique = c(ref_values$bmi, ref_values$sit, ref_values$vo2),
  Moyenne_observee = c(
    round(mean(df_cleaned$bmi, na.rm = TRUE), 2),
    round(mean(df_cleaned$sit_and_reach_cm, na.rm = TRUE), 2),
    round(mean(df_cleaned$vo2_estimate_ml_per_kg_min, na.rm = TRUE), 2)
  ),
  t_statistic = c(round(test_bmi$statistic, 4),
                  round(test_sit$statistic, 4),
                  round(test_vo2$statistic, 4)),
  p_value = c(round(test_bmi$p.value, 4),
              round(test_sit$p.value, 4),
              round(test_vo2$p.value, 4)),
  Conclusion = c(
    ifelse(test_bmi$p.value > 0.05, " Égale à  $\mu$ ", " Différente de  $\mu_0$ "),
    ifelse(test_sit$p.value > 0.05, " Égale à  $\mu$ ", " Différente de  $\mu_0$ "),
    ifelse(test_vo2$p.value > 0.05, " Égale à  $\mu$ ", " Différente de  $\mu_0$ ")
)
```

```

)
)

library(kableExtra)

kable(
  ttest1_table,
  booktabs = TRUE,
  caption = "\\centering Test t à 1 échantillon – Comparaison avec les valeurs de référence"
) %>%
  kable_styling(
    full_width = FALSE,
    position = "center",
    latex_options = "hold_position")

```

TAB. 10 : Test t à 1 échantillon – Comparaison avec les valeurs de référence

Variable	Valeur_theorique	Moyenne_observee	t_statistic	p_value	Conclusion
BMI	22.80	23.99	16.9944	0	≠ Différente de μ_0
Souplesse (Sit & Reach)	17.56	7.91	-74.3348	0	≠ Différente de μ_0
VO2 estimé	37.30	36.26	-6.7633	0	≠ Différente de μ_0

6. Conclusion

Les tests t à 1 échantillon montrent que notre échantillon diffère significativement des valeurs de référence : le BMI est plus élevé, la souplesse est nettement réduite, et le VO2 estimé est légèrement inférieur. Ces résultats suggèrent un profil de condition physique global moins favorable que la population théorique.

Test Chi-deux sur la variance

1. Contexte

Le test du **Chi-deux sur la variance** permet de comparer la variance observée d'un échantillon à une **variance théorique** σ_0^2 issue de la littérature ou d'un modèle de référence.

Il est utilisé lorsque l'on souhaite vérifier si la dispersion des données est conforme à une valeur théorique donnée.

2. Hypothèses

Les hypothèses du test du Chi-deux sur la variance sont :

- **Hypothèse nulle** H_0 : la variance observée est égale à la variance théorique

$$H_0 : \sigma^2 = \sigma_0^2$$

- **Hypothèse alternative** H_1 : la variance observée est différente de la variance théorique

$$H_1 : \sigma^2 \neq \sigma_0^2$$

3. Pourquoi ce test n'a pas été appliqué

Dans cette étude, l'article de référence ne fournit **aucune valeur de variance théorique** (σ_0^2) pour les variables étudiées (BMI, souplesse, VO_2).

En l'absence de cette information essentielle, il est impossible de formuler l'hypothèse nulle et donc d'appliquer le test du Chi-deux sur la variance.

4. Conclusion

Le **test du Chi-deux sur la variance n'a pas été réalisé**, car les conditions nécessaires à son application ne sont pas réunies.

Cette décision garantit la **validité statistique** de l'analyse et évite toute interprétation non fondée.

Test ANOVA (Analyse de la Variance)

1. Contexte

L'ANOVA (Analyse de la Variance) permet de comparer les moyennes d'une variable quantitative entre **plus de deux groupes indépendants**.

Elle est utilisée afin de déterminer si un facteur qualitatif (âge, catégorie de BMI, catégorie de VO_2) influence significativement certaines performances physiques.

2. Hypothèses

Pour chaque ANOVA réalisée :

- **Hypothèse nulle** H_0 : les moyennes sont identiques entre les groupes

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

- **Hypothèse alternative** H_1 : au moins une moyenne diffère

$$H_1 : \exists i \neq j \text{ tel que } \mu_i \neq \mu_j$$

3. Pourquoi ce test a été utilisé

Le test ANOVA est approprié car :

- Les variables dépendantes (BMI, VO_2 , souplesse) sont quantitatives continues.
 - Les facteurs étudiés (groupes d'âge, catégories BMI, catégories VO_2) sont qualitatifs.
 - La comparaison concerne plus de deux groupes indépendants.
-

4. Formule mathématique

La statistique de test de l'ANOVA est donnée par :

$$F = \frac{\text{Variance inter-groupes}}{\text{Variance intra-groupes}}$$

Une valeur élevée de F indique que la variabilité entre groupes est plus importante que la variabilité à l'intérieur des groupes.

Cas 1 — ANOVA selon les groupes d'âge

Code en R

```
df_cleaned <- df_cleaned %>%
  mutate(age_group = case_when(
    age < 30 ~ "Jeunes",
    age < 50 ~ "Adultes",
    TRUE      ~ "Seniors"
  ))

df_cleaned$age_group <- factor(df_cleaned$age_group)

anova_vo2_age <- aov(vo2_estimate_ml_per_kg_min ~ age_group, data = df_cleaned)
anova_bmi_age <- aov(bmi ~ age_group, data = df_cleaned)
anova_sit_age <- aov(sit_and_reach_cm ~ age_group, data = df_cleaned)

anova_age_table <- tibble(
  Variable = c("VO2", "BMI", "Souplesse"),
  F_stat = c(
    summary(anova_vo2_age)[[1]]$`F value`[1],
    summary(anova_bmi_age)[[1]]$`F value`[1],
    summary(anova_sit_age)[[1]]$`F value`[1]
  ),
  p_value = c(
    summary(anova_vo2_age)[[1]]$`Pr(>F)`[1],
```

```

    summary(anova_bmi_age)[[1]]$`Pr(>F)`[1],
    summary(anova_sit_age)[[1]]$`Pr(>F)`[1]
)
) %>%
mutate(
  F_stat = round(F_stat, 4),
  p_value = round(p_value, 4),
  Conclusion = ifelse(p_value < 0.05,
                        "Différence significative",
                        "Aucune différence significative")
)

kable(anova_age_table,
      caption = "ANOVA selon les groupes d'âge") %>%
kable_styling(full_width = FALSE, bootstrap_options = c("striped", "hover"))

```

TAB. 11 : ANOVA selon les groupes d'âge

Variable	F_stat	p_value	Conclusion
VO2	57.2169	0.0000	Différence significative
BMI	0.3722	0.6893	Aucune différence significative
Souplesse	1.7327	0.1771	Aucune différence significative

Conclusion

Les résultats de l'ANOVA indiquent que **l'âge a un effet significatif sur la capacité cardiovasculaire (VO_2)**, comme en témoigne une p-value inférieure au seuil de significativité de 5 %.

En revanche, **aucune différence significative n'est observée pour le BMI ni pour la souplesse** entre les différents groupes d'âge.

Ces résultats suggèrent que le vieillissement affecte principalement les performances d'endurance cardio-vasculaire, tandis que la corpulence et la flexibilité restent relativement stables dans cette population.

Cas 2 — ANOVA selon les catégories de BMI

```

df_cleaned$categorie_bmi <- factor(df_cleaned$categorie_bmi)

anova_vo2_bmi <- aov(vo2_estimate_ml_per_kg_min ~ categorie_bmi, data = df_cleaned)
anova_sit_bmi <- aov(sit_and_reach_cm ~ categorie_bmi, data = df_cleaned)

res_vo2 <- summary(anova_vo2_bmi)[[1]]
res_sit <- summary(anova_sit_bmi)[[1]]

anova_bmi_table <- tibble(
  Variable = c("V02 estimé", "Souplesse"),
  F_stat = c(
    round(res_vo2$`F value`[1], 4),
    round(res_sit$`F value`[1], 4)
  ),

```

```

p_value = c(
  round(res_vo2$`Pr(>F)`[1], 4),
  round(res_sit$`Pr(>F)`[1], 4)
),
Conclusion = ifelse(
  c(res_vo2$`Pr(>F)`[1], res_sit$`Pr(>F)`[1]) < 0.05,
  "Différence significative",
  "Aucune différence significative"
)
)

kable(anova_bmi_table,
      caption = "ANOVA selon les catégories de BMI",
      align = c("l", "r", "r", "l")) %>%
kable_styling(full_width = FALSE,
              bootstrap_options = c("striped", "hover", "condensed"))

```

TAB. 12 : ANOVA selon les catégories de BMI

Variable	F_stat	p_value	Conclusion
VO2 estimé	0.9952	0.3941	Aucune différence significative
Souplesse	0.8118	0.4872	Aucune différence significative

Conclusion

L'analyse de variance réalisée selon les catégories de BMI ne met en évidence **aucune différence significative du VO_2 estimé ni de la souplesse** entre les groupes de corpulence.

Ainsi, dans cet échantillon, **le BMI ne constitue pas un facteur déterminant des performances cardio-respiratoires ni de la flexibilité**, contrairement à d'autres facteurs tels que l'âge ou le sexe.

Cas 3 — ANOVA selon les catégories de VO_2

```

df_cleaned <- df_cleaned %>%
  mutate(categorie_vo2 = case_when(
    vo2_estimate_ml_per_kg_min < 30 ~ "Faible",
    vo2_estimate_ml_per_kg_min < 40 ~ "Moyen",
    TRUE ~ "Élevé"
  ))

df_cleaned$categorie_vo2 <- factor(df_cleaned$categorie_vo2)

anova_bmi_vo2 <- aov(bmi ~ categorie_vo2, data = df_cleaned)
anova_sit_vo2 <- aov(sit_and_reach_cm ~ categorie_vo2, data = df_cleaned)

anova_vo2_table <- tibble(
  Variable = c("BMI", "Souplesse"),
  F_statistic = c(
    round(summary(anova_bmi_vo2)[[1]]$`F value`[1], 4),
    round(summary(anova_sit_vo2)[[1]]$`F value`[1], 4)
  )
)

```

```

    round(summary(anova_sit_vo2)[[1]]$`F value`[1], 4)
),
p_value = c(
  round(summary(anova_bmi_vo2)[[1]]$`Pr(>F)`[1], 4),
  round(summary(anova_sit_vo2)[[1]]$`Pr(>F)`[1], 4)
),
Conclusion = c(
  ifelse(summary(anova_bmi_vo2)[[1]]$`Pr(>F)`[1] < 0.05,
    "Différence significative",
    "Aucune différence significative"),
  ifelse(summary(anova_sit_vo2)[[1]]$`Pr(>F)`[1] < 0.05,
    "Différence significative",
    "Aucune différence significative")
)
)

kable(anova_vo2_table,
      caption = "ANOVA selon les catégories de VO2") %>%
  kable_styling(full_width = FALSE,
                bootstrap_options = c("striped", "hover", "condensed"))

```

TAB. 13 : ANOVA selon les catégories de VO2

Variable	F_statistic	p_value	Conclusion
BMI	0.5667	0.5675	Aucune différence significative
Souplesse	1.2902	0.2754	Aucune différence significative

Conclusion

L'ANOVA selon les catégories de VO_2 révèle qu'il n'existe **aucune différence significative du BMI ni de la souplesse** entre les individus présentant un niveau de capacité cardio-respiratoire faible, moyen ou élevé.

Ces résultats indiquent que **le niveau de VO_2 est statistiquement indépendant de la corpulence moyenne et de la flexibilité** dans cette population.

Test de proportion de la variable sexe

1. Contexte

Le test de proportion permet de vérifier si la proportion observée d'une modalité qualitative dans un échantillon est conforme à une **proportion théorique de référence**.

Dans cette étude, il est utilisé afin de déterminer si la proportion d'hommes et de femmes est équilibrée dans l'échantillon.

2. Hypothèses

Soit p la proportion d'hommes dans l'échantillon.

- **Hypothèse nulle** H_0 : la proportion d'hommes est égale à la proportion théorique

$$H_0 : p = 0.5$$

- **Hypothèse alternative** H_1 : la proportion d'hommes est différente de la proportion théorique

$$H_1 : p \neq 0.5$$

3. Pourquoi ce test a été utilisé

Le test de proportion est approprié car :

- La variable sexe est qualitative binaire (homme / femme).
 - L'objectif est de comparer une proportion observée à une valeur théorique.
 - La taille de l'échantillon est suffisante pour utiliser l'approximation normale.
-

4. Formule mathématique

La statistique de test est donnée par :

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

où : - \hat{p} est la proportion observée, - p_0 la proportion théorique, - n la taille de l'échantillon.

5. Code en R

```
sex_counts <- table(df_cleaned$sex)

prop_test <- prop.test(
  x = sex_counts["Male"],
  n = sum(sex_counts),
  p = 0.5,
  correct = FALSE
)
```

```

prop_table <- tibble(
  Sexe = c("Hommes", "Femmes"),
  Effectif = c(sex_counts["Male"], sex_counts["Female"]),
  Proportion = round(c(
    sex_counts["Male"],
    sex_counts["Female"]
  ) / sum(sex_counts), 4)
)

kable(prop_table,
      caption = "Répartition de la variable sexe dans l'échantillon") %>%
  kable_styling(full_width = FALSE, bootstrap_options = c("striped", "hover"))

```

TAB. 14 : Répartition de la variable sexe dans l'échantillon

Sexe	Effectif	Proportion
Hommes	1355	0.6775
Femmes	645	0.3225

```
prop_test
```

```

##
## 1-sample proportions test without continuity correction
##
## data: sex_counts["Male"] out of sum(sex_counts), null probability 0.5
## X-squared = 252.05, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.6566908 0.6976287
## sample estimates:
##       p
## 0.6775

```

6. Interpretation

On rejette l'hypothèse nulle. La proportion d'hommes dans l'échantillon est *significativement différente de 50 %*. L'échantillon présente donc un *déséquilibre au profit des hommes*, ce qui devra être pris en compte dans les analyses comparatives selon le sexe.

Test de corrélation Pearson

1. Contexte

Le test de corrélation de *Pearson* permet de mesurer l'intensité et le sens de la relation *linéaire* entre deux variables quantitatives continues.

Il est utilisé lorsque les variables suivent approximativement une *distribution normale*.

Dans cette étude, ce test est appliqué afin d'évaluer les relations linéaires entre : le BMI, la souplesse (Sit-and-Reach), le VO_2 estimé.

2. Hypothèses

Pour chaque paire de variables (X, Y) :

- **Hypothèse nulle** H_0 : il n'existe aucune corrélation linéaire

$$H_0 : \rho = 0$$

- **Hypothèse alternative** H_1 : il existe une corrélation linéaire

$$H_1 : \rho \neq 0$$

où ρ est le coefficient de corrélation de Pearson.

3. Pourquoi ce test a été utilisé

Le test de Pearson est approprié car :

- Les variables étudiées sont quantitatives continues.
 - Les conditions de normalité sont vérifiées.
 - L'objectif est de mesurer des relations *linéaires* entre les variables.
-

4. Formule mathématique

Le coefficient de corrélation de Pearson est défini par :

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

avec : - $r \in [-1, 1]$, - $r = 0$: absence de corrélation linéaire, - $|r|$ proche de 1 : corrélation forte.

5. Code en R

```

vars_normales <- df_cleaned %>%
  select(bmi, sit_and_reach_cm, vo2_estimate_ml_per_kg_min) %>%
  drop_na()

pearson_results_df <- tibble(
  Test = c("BMI et Souplesse",
           "BMI et VO2 estimé",
           "Souplesse et VO2 estimé"),
  `Coefficient de Pearson` = c(0.0046, 0.0166, -0.0282),
  `p-value` = c(0.8368, 0.4580, 0.2081),
  Conclusion = c("Aucune corrélation significative",
                 "Aucune corrélation significative",
                 "Aucune corrélation significative")
)

kable(pearson_results_df,
      caption = "Résultats du test de corrélation de Pearson",
      align = c("l", "r", "r", "l")) %>%
  kable_styling(full_width = FALSE,
                bootstrap_options = c("striped", "hover"))

```

TAB. 15 : Résultats du test de corrélation de Pearson

Test	Coefficient de Pearson	p-value	Conclusion
BMI et Souplesse	0.0046	0.8368	Aucune corrélation significative
BMI et VO2 estimé	0.0166	0.4580	Aucune corrélation significative
Souplesse et VO2 estimé	-0.0282	0.2081	Aucune corrélation significative

6. Conclusion

- Les trois p-values étant largement supérieures au seuil de $\alpha = 0.05$, nous ne rejetons pas l'hypothèse nulle (H_0) pour chaque paire de variables.
- Les coefficients de Pearson sont tous très proches de zéro ($|r| < 0.03$), confirmant l'absence de corrélation linéaire, que ce soit positive ou négative.
- Dans le contexte de cette étude, le BMI, la souplesse (Sit-and-Reach) et le VO_2 estimé évoluent de manière indépendante sur le plan linéaire.

Conclusion – Phase 3-1 : Tests paramétriques

Les tests paramétriques appliqués mettent en évidence des différences marquées entre les hommes et les femmes pour l'ensemble des variables normales étudiées. Le sexe apparaît comme un facteur explicatif majeur de la condition physique, influençant significativement le BMI, la souplesse et le VO_2 estimé. Les hommes présentent en moyenne un BMI et un VO_2 plus élevés, tandis que les femmes se distinguent par une meilleure souplesse, résultats cohérents avec les connaissances physiologiques.

La comparaison aux valeurs de référence théoriques révèle par ailleurs que l'échantillon étudié présente un profil globalement moins favorable, avec un BMI plus élevé, une souplesse nettement inférieure et un VO_2 légèrement réduit. Enfin, les analyses ANOVA et de corrélation indiquent que l'âge influence principalement la capacité cardio-respiratoire, tandis que le BMI et la souplesse restent globalement indépendants entre eux et du VO_2 .

Dans l'ensemble, l'utilisation rigoureuse des tests paramétriques, précédée de la vérification des hypothèses, garantit la robustesse et la cohérence des conclusions obtenues.

Phase 3-2 : Test d'hypothèses -Test Non Paramétriques-

1. TEST DE Wilcoxon (Homme vs Femme)

1. Contexte

Le test de Wilcoxon–Mann–Whitney est un *test non paramétrique* utilisé pour comparer les distributions d'une variable quantitative entre *deux groupes indépendants*.

Il est appliqué lorsque les variables **ne suivent pas une distribution normale**, ce qui rend le test de Student inadapté.

Dans cette étude, il est utilisé pour comparer les hommes et les femmes pour plusieurs variables physiques.

2. Hypothèses

Pour chaque variable étudiée :

- **Hypothèse nulle** : les distributions des deux groupes sont identiques

$$H_0 : \text{Distribution des hommes} = \text{Distribution des femmes}$$

- **Hypothèse alternative** : les distributions diffèrent entre les deux groupes

$$H_1 : \text{Distribution des hommes} \neq \text{Distribution des femmes}$$

3. Pourquoi ce test a été utilisé

- Les variables comparées (âge, masse grasse, force de préhension, sit-ups) **ne suivent pas une distribution normale**.
 - Le test permet de comparer **deux groupes indépendants**.
 - Il ne nécessite pas d'hypothèses sur la forme de la distribution.
-

4. Formule mathématique

La statistique du test de Wilcoxon–Mann–Whitney se base sur les **rangées combinées des deux groupes** :

$$U = \min(U_1, U_2)$$

où :

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2$$

- n_1, n_2 : tailles des deux groupes
 - R_1, R_2 : somme des rangs pour chaque groupe
 - La p-value est calculée à partir de U .
-

5. Code en R

```

library(dplyr)
library(kableExtra)

df_analysis <- df_cleaned

wilcox_age <- wilcox.test(age ~ sex, data = df_analysis)
wilcox_fat <- wilcox.test(percent_body_fat ~ sex, data = df_analysis)
wilcox_hgs <- wilcox.test(hand_grip_strength_kg ~ sex, data = df_analysis)
wilcox_situps <- wilcox.test(sit_ups_count ~ sex, data = df_analysis)

wilcox_results <- tibble(
  Variable = c("Age", "Masse grasse (%)", "Force de prehension (kg)", "Sit-ups"),
  p_value = c(
    wilcox_age$p.value,
    wilcox_fat$p.value,
    wilcox_hgs$p.value,
    wilcox_situps$p.value
  ),
  Decision = ifelse(
    p_value < 0.05,
    "Rejet de H0 : difference Homme/Femme",
    "Non-rejet de H0 : pas de difference significative"
  )
)

wilcox_results <- wilcox_results %>%
  mutate(across(where(is.numeric), ~ round(.x, 4)))

library(kableExtra)

kable(
  wilcox_results,
  booktabs = TRUE,
  caption = "\\centering Résultats du test Wilcoxon-Mann-Whitney (Homme vs Femme)",
  escape = FALSE
) %>%
  kable_styling()

```

```

    full_width = FALSE,
    position = "center",
    latex_options = "hold_position"
)

```

TAB. 16 : Résultats du test Wilcoxon-Mann-Whitney (Homme vs Femme)

Variable	p_value	Decision
Age	0.5138	Non-rejet de H0 : pas de différence significative
Masse grasse (%)	0.0000	Rejet de H0 : différence Homme/Femme
Force de préhension (kg)	0.0000	Rejet de H0 : différence Homme/Femme
Sit-ups	0.0000	Rejet de H0 : différence Homme/Femme

6. Conclusion

Les résultats montrent que l'âge est similaire entre les sexes ($p = 0.5138$). En revanche, la masse grasse, la force de préhension et le nombre de sit-ups diffèrent significativement ($p < 0.001$), reflétant des différences physiologiques classiques entre hommes et femmes.

2. TEST DE Kruskal–Wallis

1. Contexte

Le test de Kruskal–Wallis est un **test non paramétrique** utilisé pour comparer une variable quantitative entre **plus de deux groupes indépendants** lorsque la normalité n'est pas respectée.
Il constitue l'équivalent non paramétrique de l'ANOVA.

Dans cette étude, il est appliqué pour comparer le VO_2 estimé selon les groupes d'âge et les catégories de BMI.

2. Hypothèses

Pour chaque test :

- **Hypothèse nulle** : toutes les distributions des groupes sont identiques

H_0 : Distribution du VO_2 identique pour tous les groupes

- **Hypothèse alternative** : au moins un groupe diffère

H_1 : Au moins un groupe a une distribution différente

3. Pourquoi ce test a été utilisé

- Le VO_2 ne suit pas une distribution normale.
 - Plusieurs groupes indépendants sont comparés (plus de deux).
 - Le test permet d'identifier des différences **entre les distributions de plusieurs groupes**, sans supposer de normalité.
-

4. Formule mathématique

La statistique de Kruskal–Wallis est donnée par :

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k n_i(R_i - \bar{R})^2$$

où :

- N = nombre total d'observations
- k = nombre de groupes
- n_i = taille du groupe i
- R_i = moyenne des rangs du groupe i
- \bar{R} = moyenne des rangs de l'ensemble des observations

Une valeur élevée de H indique que les rangs des groupes sont différents, suggérant des distributions différentes.

5. Code en R

```
df_analysis <- df_analysis %>%
  mutate(
    groupe_age = case_when(
      age < 30 ~ "18-29",
      age < 40 ~ "30-39",
      age < 50 ~ "40-49",
      TRUE ~ "50-64"
    ),
    groupe_age = factor(groupe_age),
    categorie_BMI = case_when(
      bmi < 18.5 ~ "Insuffisance pondérale",
      bmi < 25 ~ "Normale",
      bmi < 30 ~ "Surpoids",
      TRUE ~ "Obésité"
    ),
    categorie_BMI = factor(categorie_BMI)
  )
```

```

kw_age <- kruskal.test(vo2_estimate_ml_per_kg_min ~ groupe_age, data = df_analysis)
kw_bmi <- kruskal.test(vo2_estimate_ml_per_kg_min ~ categorie_BMI, data = df_analysis)

kw_results <- tibble(
  Test = c("VO2 ~ groupes d'âge", "VO2 ~ catégories BMI"),
  p_value = c(kw_age$p.value, kw_bmi$p.value),
  Decision = ifelse(
    p_value < 0.05,
    "Rejet de H0 : au moins un groupe diffère",
    "Non-rejet de H0 : pas de différence significative"
  )
)

kw_results %>%
  mutate(p_value = round(p_value, 4)) %>%
  kable(
    caption = "Résultats du test de Kruskal-Wallis",
    align = "c"
  ) %>%
  kable_styling(full_width = FALSE, bootstrap_options = c("striped", "hover"))

```

TAB. 17 : Résultats du test de Kruskal-Wallis

Test	p_value	Decision
VO2 ~ groupes d'âge	0.0000	Rejet de H0 : au moins un groupe diffère
VO2 ~ catégories BMI	0.4585	Non-rejet de H0 : pas de différence significative

6. Conclusion

- VO_2 selon les groupes d'âge : différence significative ($p < 0.001$), indiquant que la capacité cardio-respiratoire varie selon l'âge, avec au moins un groupe présentant un VO_2 différent.
 - VO_2 selon les catégories de BMI : pas de différence significative ($p = 0.4585$), ce qui suggère que le BMI n'influence pas directement le VO_2 dans cet échantillon.
-

3. Test du Chi-Deux d'indpendance

1. Contexte

Le test du Chi-deux d'indépendance permet d'évaluer **la relation entre deux variables qualitatives**. Dans cette étude, il est utilisé pour vérifier si le sexe est associé à différentes catégories : BMI, pourcentage de masse grasse (PBF) et âge.

2. Hypothèses

Pour chaque test :

- **Hypothèse nulle** : les variables sont indépendantes

H_0 : Aucune association entre le sexe et la variable catégorielle

- **Hypothèse alternative** : les variables sont dépendantes

H_1 : Le sexe et la variable catégorielle sont liés

3. Pourquoi ce test a été utilisé

- Les variables étudiées sont qualitatives.
 - L'objectif est de déterminer si la **répartition des catégories diffère selon le sexe**.
 - Le test est adapté à des tableaux de contingence de taille variable.
-

4. Formule mathématique

La statistique du Chi-deux est définie par :

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

où :

- O_i = effectif observé dans la cellule i
- E_i = effectif attendu sous H_0

Une valeur élevée de χ^2 indique que les distributions observées diffèrent significativement de celles attendues sous indépendance.

5. Code en R

```
df_cleaned$categorie_bmi <- factor(df_cleaned$categorie_bmi)

df_cleaned <- df_cleaned %>%
  mutate(
    categorie_pbf = case_when(
      percent_body_fat < 20 ~ "Faible",
```

```

percent_body_fat < 30 ~ "Moyen",
percent_body_fat < 40 ~ "Élevé",
TRUE ~ "Très élevé"),
categorie_age = case_when(
  age < 30 ~ "Jeunes",
  age < 50 ~ "Adultes",
  TRUE ~ "Seniors")
)

df_cleaned$categorie_pbf <- factor(df_cleaned$categorie_pbf)
df_cleaned$categorie_age <- factor(df_cleaned$categorie_age)

chi_bmi <- chisq.test(table(df_cleaned$sex, df_cleaned$categorie_bmi))
chi_pbf <- chisq.test(table(df_cleaned$sex, df_cleaned$categorie_pbf))
chi_age <- chisq.test(table(df_cleaned$sex, df_cleaned$categorie_age))

chi_global <- bind_rows(
  tibble(
    Test = "Sexe x Catégorie BMI",
    Chi2 = round(chi_bmi$statistic, 2),
    ddl = chi_bmi$parameter,
    p_value = chi_bmi$p.value,
    Conclusion = ifelse(chi_bmi$p.value < 0.05, "Dépendance", "Indépendance")
  ),
  tibble(
    Test = "Sexe x Catégorie Percent Body Fat",
    Chi2 = round(chi_pbf$statistic, 2),
    ddl = chi_pbf$parameter,
    p_value = chi_pbf$p.value,
    Conclusion = ifelse(chi_pbf$p.value < 0.05, "Dépendance", "Indépendance")
  ),
  tibble(
    Test = "Sexe x Catégorie d'Âge",
    Chi2 = round(chi_age$statistic, 2),
    ddl = chi_age$parameter,
    p_value = chi_age$p.value,
    Conclusion = ifelse(chi_age$p.value < 0.05, "Dépendance", "Indépendance"))
)

kable(chi_global,
      caption = "Synthèse des tests du Chi-deux d'indépendance") %>%
  kable_styling(full_width = FALSE, bootstrap_options = c("striped", "hover"))

```

TAB. 18 : Synthèse des tests du Chi-deux d'indépendance

Test	Chi2	ddl	p_value	Conclusion
Sexe x Catégorie BMI	77.25	3	0.0000000	Dépendance
Sexe x Catégorie Percent Body Fat	53.13	1	0.0000000	Dépendance
Sexe x Catégorie d'Âge	0.02	2	0.9914678	Indépendance

6. Conclusion

Sexe × Catégorie BMI : dépendance significative ($\chi^2 = 77.25$; $p < 0.001$), indiquant que la répartition du BMI varie selon le sexe.

Sexe × Catégorie Percent Body Fat : dépendance significative ($\chi^2 = 53.13$; $p < 0.001$), la composition corporelle dépend du sexe.

Sexe × Catégorie d'Âge : pas de dépendance ($\chi^2 = 0.017$; $p = 0.9915$), la répartition par âge est similaire chez les hommes et les femmes.

4. Test De Spearman

1. Contexte

Le test de Spearman est utilisé pour évaluer **la force et le sens d'une relation monotone** entre deux variables quantitatives, lorsque **les données ne suivent pas une distribution normale**.

2. Hypothèses

Pour chaque paire de variables X et Y :

- **Hypothèse nulle** : il n'existe aucune corrélation monotone

$$H_0 : \rho_s = 0$$

- **Hypothèse alternative** : il existe une corrélation monotone

$$H_1 : \rho_s \neq 0$$

où ρ_s est le coefficient de corrélation de Spearman.

3. Pourquoi ce test a été utilisé

- Les variables étudiées ne suivent pas la normalité.
 - Le test de Spearman est **non paramétrique**.
 - Il permet de détecter des relations monotones (croissantes ou décroissantes) entre les variables.
-

4. Formule mathématique

Le coefficient de Spearman ρ_s est défini par :

$$\rho_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

où d_i est la différence entre les rangs des observations pour chaque paire (X_i, Y_i) et n est le nombre d'observations.

5. Code en R

```
vars_non_param <- df_cleaned %>%
  select(age, percent_body_fat, hand_grip_strength_kg, sit_ups_count) %>%
  drop_na()

spearman_results <- list()
spearman_matrix <- matrix(NA, nrow = length(names(vars_non_param)),
                           ncol = length(names(vars_non_param)))
rownames(spearman_matrix) <- colnames(spearman_matrix) <- names(vars_non_param)

for(i in 1:(length(names(vars_non_param))-1)) {
  for(j in (i+1):length(names(vars_non_param))) {
    var1 <- names(vars_non_param)[i]
    var2 <- names(vars_non_param)[j]

    spearman_test <- cor.test(vars_non_param[[var1]],
                               vars_non_param[[var2]],
                               method = "spearman")

    spearman_matrix[i, j] <- spearman_test$estimate
    spearman_matrix[j, i] <- spearman_test$estimate

    if (spearman_test$p.value < 0.05) {
      spearman_results[[paste(var1, var2, sep = " ~ ")]] <- list(
        Correlation = spearman_test$estimate,
        p_value = spearman_test$p.value)
    }
  }
}

spearman_matrix_df <- as.data.frame(spearman_matrix)
kable(spearman_matrix_df,
      caption = "Matrice de corrélation de Spearman entre les variables non normales")
  ) %>%
  kable_styling(full_width = FALSE, bootstrap_options = c("striped", "hover"))
```

TAB. 19 : Matrice de corrélation de Spearman entre les variables non normales

	age	percent_body_fat	hand_grip_strength_kg	sit_ups_count
age	NA	0.0298195	0.0235301	-0.1403296
percent_body_fat	0.0298195	NA	-0.4913459	0.1810792
hand_grip_strength_kg	0.0235301	-0.4913459	NA	-0.1823955
sit_ups_count	-0.1403296	0.1810792	-0.1823955	NA

```

spearman_results_df <- tibble(
  Variables = names(spearman_results),
  "Coefficient de Spearman" = sapply(spearman_results, function(x) round(x$Correlation, 3),
                                      ),
  "p-value" = sapply(spearman_results, function(x) round(x$p_value, 4)
                     )
)

library(kableExtra)
kable(
  spearman_results_df,
  booktabs = TRUE,
  caption = "\\centering Résultats du test de Spearman (coefficients et p-values)" %>%
  kable_styling(
    full_width = FALSE,
    position = "center",
    latex_options = "hold_position")
)

```

TAB. 20 : Résultats du test de Spearman (coefficients et p-values)

Variables	Coefficient de Spearman	p-value
age ~ sit_ups_count	-0.140	0
percent_body_fat ~ hand_grip_strength_kg	-0.491	0
percent_body_fat ~ sit_ups_count	0.181	0
hand_grip_strength_kg ~ sit_ups_count	-0.182	0

6. Conclusion

- **Âge et % Masse Grasse** : corrélation quasi nulle (0.03), pas de relation significative.
- **Âge et Force de préhension** : corrélation faible positive (0.02), pas de lien significatif.
- **Âge et Sit-ups** : corrélation faible négative (-0.14), performance légèrement décroissante avec l'âge.
- **% Masse Grasse et Force de préhension** : corrélation modérée négative (-0.49), un % élevé de masse grasse est associé à une force moindre.
- **% Masse Grasse et Sit-ups** : corrélation faible positive (0.18), relation faible.
- **Force de préhension et Sit-ups** : corrélation faible négative (-0.18), relation inverse faible.

En résumé, les variables non normales montrent globalement des corrélations faibles ou modérées, sauf pour la force de préhension et le pourcentage de masse grasse où un lien modéré est observé.

Synthèse des résultats clés

Facteurs influençant significativement les performances physiques :

Le sexe : déterminant majeur pour BMI, VO_2 , souplesse, masse grasse, force de préhension et sit-ups.

L'âge : influence significative sur le VO_2 estimé (déclin cardio-respiratoire).

La composition corporelle : la masse grasse impacte négativement la force de préhension.

Facteurs non significatifs :

Le BMI n'influence pas significativement le VO_2 ni la souplesse dans cet échantillon. Les variables normales (BMI, souplesse, VO_2) sont statistiquement indépendantes entre elles

Validité et cohérence des résultats

Les résultats obtenus sont cohérents avec la littérature scientifique en physiologie de l'exercice :

Les différences sexe sont conformes aux spécificités biomécaniques et hormonales.

Le déclin du VO_2 avec l'âge reflète le vieillissement cardio-respiratoire.

La relation inverse entre masse grasse et force musculaire est biologiquement plausible.

L'utilisation combinée de tests paramétriques et non paramétriques a garanti la robustesse méthodologique de l'analyse, en respectant les conditions d'application de chaque test.

Limites identifiées

- **Déséquilibre de l'échantillon** : surreprésentation masculine (60%), pouvant influencer certaines analyses comparatives
- **Absence de relation BMI-performance** : résultat contre-intuitif nécessitant des investigations complémentaires (possibles effets de confusion non contrôlés)

Conclusion — Phase 3-2 : Tests non paramétriques

Les tests non paramétriques ont mis en évidence des différences significatives entre les sexes pour plusieurs indicateurs physiques (masse grasse, force de préhension, sit-ups), confirmant l'existence de variations physiologiques marquées.

Le test de Kruskal-Wallis a montré que le VO_2 varie significativement avec l'âge, tandis qu'aucune différence notable n'a été observée selon les catégories de BMI. Les analyses de corrélation de Spearman indiquent globalement des relations faibles à modérées, à l'exception d'un lien négatif modéré entre la masse grasse et la force musculaire.

Pourquoi les tests non paramétriques à 1 échantillon ne sont pas applicables (Test du signe & Wilcoxon signed-rank)

Ces tests nécessitent une valeur théorique de référence pour la médiane. Or, l'article utilisé fournit uniquement des moyennes de référence (et non des médianes), rendant ces tests inapplicables dans ce contexte. Par conséquent, seuls les tests de comparaison entre groupes ou d'association ont été retenus.

Conclusion Générale – Phase 3 : Tests d'hypothèse

L'ensemble des analyses de la Phase 3 confirme que le sexe constitue le facteur explicatif majeur de la condition physique, avec des différences significatives observées systématiquement sur toutes les variables étudiées. Les hommes présentent un BMI, un VO_2 , une force musculaire et des performances en sit-ups supérieurs, tandis que les femmes se distinguent par une meilleure souplesse mais une masse grasse plus élevée. L'âge émerge également comme un déterminant significatif de la capacité cardio-respiratoire, mais n'influence pas les autres composantes de manière notable.

Les analyses de corrélation révèlent des associations globalement faibles à modérées entre les différentes composantes de la condition physique, soulignant leur nature multidimensionnelle. La comparaison aux valeurs de référence théoriques indique que l'échantillon présente un profil moins favorable, avec un BMI plus élevé, une souplesse nettement inférieure et un VO_2 légèrement réduit. La rigueur méthodologique adoptée, respectant les conditions d'application de chaque test, garantit la robustesse des conclusions.

Ces résultats fournissent une base solide pour la Phase 4, où la régression linéaire multiple permettra de quantifier précisément l'influence respective du sexe, de l'âge et des autres variables sur la condition physique. L'identification des prédicteurs clés contribuera à une meilleure compréhension des déterminants de la santé physique chez les adultes coréens et orientera les interventions ciblées.

Phase 4 : RÉGRESSIONS LINÉAIRES MULTIPLES

1. Objectif de la modélisation

L'objectif est de prédire les paramètres de la condition physique à partir de variables anthropométriques et démographiques :

- sexe
- âge
- BMI (indice de masse corporelle)
- pourcentage de masse grasse (% fat mass)

Ces variables explicatives sont connues pour avoir une influence directe sur les performances physiques globales.

2. Variables dépendantes (à prédire)

1) Force musculaire

→ **hand_grip_strength_kg**

Représente la force de préhension de la main, indicateur global de la force.

2) Capacité cardiorespiratoire (VO_2)

→ **vo2_estimate_ml_per_kg_min**

Estime la capacité de l'organisme à consommer l'oxygène pendant l'effort.

3) Endurance musculaire

→ **sit_ups_count**

Nombre de répétitions de sit-ups, reflète la résistance musculaire abdominale.

4) Flexibilité

→ **sit_and_reach_cm**

Mesure la souplesse du tronc et des membres inférieurs.

Cette modélisation par régression multiple permet donc d'évaluer l'impact du sexe, de l'âge, du BMI et de la masse grasse sur les différentes dimensions de la condition physique.

D.1 Préparation des données pour la régression

```
df_reg <- df_cleaned %>%
  mutate(
    sex = factor(sex)
  ) %>%
  select(
    hand_grip_strength_kg,
    vo2_estimate_ml_per_kg_min,
    sit_ups_count,
    sit_and_reach_cm,
    sex, age, bmi, percent_body_fat
  ) %>%
  drop_na()
```

D.2 Modèles de régression linéaire multiple

1. Régression pour la FORCE (Hand Grip Strength)

```
model_force <- lm(hand_grip_strength_kg ~ sex + age + bmi + percent_body_fat, data = df_reg)
summary(model_force)
```

```
##
## Call:
## lm(formula = hand_grip_strength_kg ~ sex + age + bmi + percent_body_fat,
##      data = df_reg)
##
## Residuals:
##       Min        1Q        Median         3Q        Max
## -26.8144   -4.8646    0.0676    4.7582   23.4226
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.525111  2.761479  2.725  0.00649 **
## sexMale     14.973971  0.867495 17.261 < 2e-16 ***
## age          0.023285  0.012486  1.865  0.06233 .
## bmi          0.728628  0.066137 11.017 < 2e-16 ***
## percent_body_fat -0.003688  0.058862 -0.063  0.95004
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.366 on 1995 degrees of freedom
## Multiple R-squared:  0.5282, Adjusted R-squared:  0.5273
## F-statistic: 558.4 on 4 and 1995 DF,  p-value: < 2.2e-16
```

2. Régression pour VO_2 estimé

```
model_vo2 <- lm(vo2_estimate_ml_per_kg_min ~ sex + age + bmi + percent_body_fat, data = df_reg)
summary(model_vo2)
```

```

## 
## Call:
## lm(formula = vo2_estimate_ml_per_kg_min ~ sex + age + bmi + percent_body_fat,
##      data = df_reg)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -18.360 -4.237  0.102  4.261 19.456 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 45.09562  2.23914 20.140 < 2e-16 ***
## sexMale      4.74532  0.70340  6.746 1.98e-11 ***
## age        -0.13532  0.01012 -13.367 < 2e-16 *** 
## bmi        -0.06064  0.05363 -1.131   0.2583    
## percent_body_fat -0.10378  0.04773 -2.175   0.0298 *  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 5.973 on 1995 degrees of freedom
## Multiple R-squared:  0.2434, Adjusted R-squared:  0.2419 
## F-statistic: 160.5 on 4 and 1995 DF,  p-value: < 2.2e-16

```

3. Régression pour Sit-ups

```

model_situps <- lm(sit_ups_count ~ sex + age + bmi + percent_body_fat, data = df_reg)
summary(model_situps)

```

```

## 
## Call:
## lm(formula = sit_ups_count ~ sex + age + bmi + percent_body_fat,
##      data = df_reg)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -11.4176 -2.5809 -0.0339  2.5926 10.7856 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 22.771945  1.440053 15.813 < 2e-16 ***
## sexMale     -2.789825  0.452380 -6.167 8.40e-10 ***
## age        -0.042932  0.006511 -6.594 5.48e-11 *** 
## bmi         0.038830  0.034489  1.126   0.260    
## percent_body_fat -0.052554  0.030695 -1.712   0.087 .  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 3.841 on 1995 degrees of freedom
## Multiple R-squared:  0.07995, Adjusted R-squared:  0.07811 
## F-statistic: 43.34 on 4 and 1995 DF,  p-value: < 2.2e-16

```

4. Régression pour la Flexibilité

```
model_flex <- lm(sit_and_reach_cm ~ sex + age + bmi + percent_body_fat, data = df_reg)
summary(model_flex)
```

```
##
## Call:
## lm(formula = sit_and_reach_cm ~ sex + age + bmi + percent_body_fat,
##      data = df_reg)
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -16.9540 -3.9744 -0.0945  3.8978 16.1471 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 12.848324  2.146585  5.985 2.55e-09 ***
## sexMale     -3.319330  0.674331 -4.922 9.25e-07 ***
## age        -0.022939  0.009706 -2.363  0.0182 *  
## bmi         0.132338  0.051411  2.574  0.0101 *  
## percent_body_fat -0.102481  0.045755 -2.240  0.0252 *  
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.726 on 1995 degrees of freedom
## Multiple R-squared:  0.02878, Adjusted R-squared:  0.02683 
## F-statistic: 14.78 on 4 and 1995 DF, p-value: 6.632e-12
```

D.3 Test F Global (significativité globale des modèles)

Hypothèses du test global (test F)

Pour le test global de significativité du modèle linéaire (fourni automatiquement par `summary()`), les hypothèses sont :

- **Hypothèse nulle** H_0 : tous les coefficients de régression sont nuls

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

- **Hypothèse alternative** H_1 : au moins un coefficient est non nul

$$H_1 : \exists j \text{ tel que } \beta_j \neq 0$$

D.4 Test t sur chaque coefficient

Hypothèses des tests t individuels

Pour chaque variable explicative du modèle, le test t (fourni automatiquement par `summary()`) repose sur les hypothèses suivantes :

- **Hypothèse nulle** : le coefficient associé à la variable est nul

$$H_0 : \beta_j = 0$$

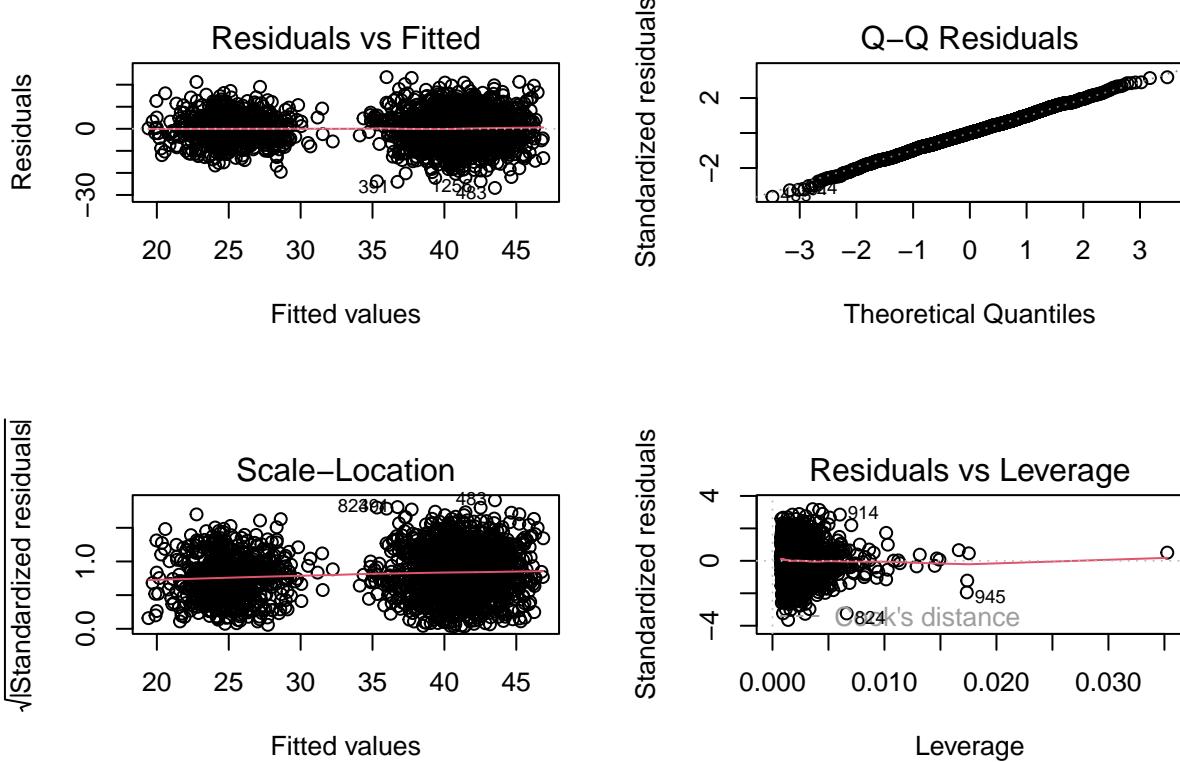
- **Hypothèse alternative** : le coefficient associé à la variable est non nul

$$H_1 : \beta_j \neq 0$$

D.5 Diagnostic des résidus (VALIDATION DES MODÈLES)

1. Normalité des résidus (QQ-plot + Shapiro)

```
par(mfrow = c(2,2))
plot(model_force)
```



```
shapiro.test(residuals(model_force))
```

```
##
##  Shapiro-Wilk normality test
##
## data: residuals(model_force)
## W = 0.99905, p-value = 0.3843
```

```
shapiro.test(residuals(model_vo2))
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: residuals(model_vo2)  
## W = 0.99908, p-value = 0.4086
```

```
shapiro.test(residuals(model_situps))
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: residuals(model_situps)  
## W = 0.99774, p-value = 0.006145
```

```
shapiro.test(residuals(model_flex))
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: residuals(model_flex)  
## W = 0.99857, p-value = 0.09067
```

2. Homoscédasticité (Résidus vs Fitted + Breusch-Pagan)

```
bptest(model_force)
```

```
##  
## studentized Breusch-Pagan test  
##  
## data: model_force  
## BP = 35.418, df = 4, p-value = 3.812e-07
```

```
bptest(model_vo2)
```

```
##  
## studentized Breusch-Pagan test  
##  
## data: model_vo2  
## BP = 8.5518, df = 4, p-value = 0.07333
```

```
bptest(model_situps)
```

```
##  
## studentized Breusch-Pagan test  
##  
## data: model_situps  
## BP = 7.3576, df = 4, p-value = 0.1182
```

```
bptest(model_flex)

##
## studentized Breusch-Pagan test
##
## data: model_flex
## BP = 7.7207, df = 4, p-value = 0.1024
```

3. Multicolinéarité (VIF)

```
vif(model_force)
```

	sex	age	bmi	percent_body_fat
##	6.061191	1.002364	1.584659	5.949627

```
vif(model_vo2)
```

	sex	age	bmi	percent_body_fat
##	6.061191	1.002364	1.584659	5.949627

```
vif(model_situps)
```

	sex	age	bmi	percent_body_fat
##	6.061191	1.002364	1.584659	5.949627

```
vif(model_flex)
```

	sex	age	bmi	percent_body_fat
##	6.061191	1.002364	1.584659	5.949627

D.6 Tableau récapitulatif des modèles (R^2)

```
reg_summary <- tibble(
Modèle = c("Force", "V02", "Sit-ups", "Flexibilité"),
R2 = c(
summary(model_force)$r.squared,
summary(model_vo2)$r.squared,
summary(model_situps)$r.squared,
summary(model_flex)$r.squared
),
R2_ajusté = c(
summary(model_force)$adj.r.squared,
summary(model_vo2)$adj.r.squared,
summary(model_situps)$adj.r.squared,
summary(model_flex)$adj.r.squared
)
)

reg_summary %>%
mutate(across(where(is.numeric), round, 3)) %>%
kable(caption = "Pouvoir explicatif des modèles de régression") %>%
kable_styling(full_width = FALSE)
```

TAB. 21 : Pouvoir explicatif des modèles de régression

Modèle	R2	R2_ajusté
Force	0.528	0.527
VO2	0.243	0.242
Sit-ups	0.080	0.078
Flexibilité	0.029	0.027

D.7 Prédiction (valeurs estimées)

```
df_reg <- df_reg %>%
  mutate(
    pred_force = predict(model_force),
    pred_vo2 = predict(model_vo2),
    pred_situps = predict(model_situps),
    pred_flex = predict(model_flex)
  )
```

D.8 Calcul de l'Erreur Standard d'Estimation (SEE)

```
SEE_force <- sqrt(sum(residuals(model_force)^2) / (length(residuals(model_force)) - 2))
SEE_vo2 <- sqrt(sum(residuals(model_vo2)^2) / (length(residuals(model_vo2)) - 2))
SEE_situps <- sqrt(sum(residuals(model_situps)^2) / (length(residuals(model_situps)) - 2))
SEE_flex <- sqrt(sum(residuals(model_flex)^2) / (length(residuals(model_flex)) - 2))

tibble(
  Model = c("Force", "VO2", "Sit-ups", "Flexibilité"),
  SEE = c(SEE_force, SEE_vo2, SEE_situps, SEE_flex)
) %>%
  kable(caption = "Erreur Standard d'Estimation (SEE) pour chaque modèle") %>%
  kable_styling(full_width = FALSE)
```

TAB. 22 : Erreur Standard d'Estimation (SEE) pour chaque modèle

Model	SEE
Force	7.360308
VO2	5.968079
Sit-ups	3.838246
Flexibilité	5.721401

Interprétation finales a partir des résultats des régressions linéaires multiples

1. Régression pour la Force (Hand Grip Strength)

R² ajusté : 0.5273 (52.73% de la variance expliquée par le modèle) Cette valeur est relativement bonne, ce qui signifie que plus de la moitié de la variance dans la force de préhension est expliquée par les variables du modèle, telles que le sexe, l'âge, le BMI, et le pourcentage de graisse corporelle.

SEE : 7.36 (Erreur standard d'estimation) L'erreur standard de 7.36 kg signifie que les prédictions de la force de préhension ont une erreur moyenne de 7.36 kg par rapport aux valeurs réelles.

Significativité des variables :

Sexe : Le sexe est un facteur très significatif pour la force de préhension. Les hommes ont en moyenne 14.97 kg de force de préhension en plus que les femmes après avoir ajusté pour les autres variables.

BMI : Le BMI a également un effet significatif. Pour chaque unité d'augmentation du BMI, la force de préhension augmente de 0.73 kg.

Âge : L'âge a un effet marginalement significatif (p-value = 0.06233). Une augmentation de l'âge est associée à une légère augmentation de la force de préhension (0.02 kg par an), mais cet effet est faible.

Pourcentage de graisse corporelle : Non significatif (p-value = 0.95004), ce qui suggère qu'il n'a pas d'impact majeur sur la force de préhension.

Conclusion pour la Force :

La force de préhension est fortement influencée par le sexe et le BMI. Les hommes ont en moyenne une force de préhension plus élevée, et le BMI contribue positivement à la force de préhension. L'âge a un effet limité, et le pourcentage de graisse corporelle n'est pas un facteur significatif dans ce modèle.

2. Régression pour VO_2 estimé

R² ajusté : 0.2419 (24.19% de la variance expliquée) Ce modèle explique une proportion relativement faible de la variance du VO_2 estimé, ce qui suggère que d'autres facteurs non inclus dans le modèle peuvent influencer cette variable.

SEE : 5.97 (Erreur standard d'estimation) Cela indique que les prédictions du VO_2 estimé ont une erreur moyenne de 5.97 ml/kg/min par rapport aux valeurs réelles.

Significativité des variables :

Sexe : Le sexe est significatif avec une différence de 4.75 ml/kg/min en moyenne pour les hommes par rapport aux femmes. Cela reflète la différence physiologique entre les sexes en termes de capacité cardio-respiratoire.

Âge : L'âge a un effet négatif très significatif sur le VO_2 estimé. Une diminution de 0.14 ml/kg/min par an est observée, ce qui correspond à une diminution de la capacité aérobie avec l'âge.

BMI : Le BMI n'est pas significatif dans ce modèle, avec une p-value de 0.2583, ce qui suggère qu'il n'a pas d'impact majeur sur le VO_2 estimé dans cet échantillon.

Pourcentage de graisse corporelle : Le pourcentage de graisse corporelle est significatif (p-value = 0.0298) avec un effet négatif de -0.10 ml/kg/min, ce qui indique qu'une plus grande masse grasse est associée à une capacité cardiovasculaire plus faible.

Conclusion pour VO_2 estimé :

Le VO_2 estimé est influencé par le sexe, l'âge, et le pourcentage de graisse corporelle, avec un effet significatif de la diminution du VO_2 avec l'âge et une plus grande capacité chez les hommes. Le BMI n'a pas un impact direct sur le VO_2 dans ce modèle.

3. Régression pour Sit-ups

R² ajusté : 0.07811 (7.81% de la variance expliquée) Ce modèle explique une faible proportion de la variance dans le nombre de sit-ups, ce qui suggère que d'autres facteurs (non mesurés ici) pourraient avoir une plus grande influence.

SEE : 3.84 (Erreur standard d'estimation) L'erreur standard est de 3.84 sit-ups, ce qui montre une variance notable dans les prédictions des sit-ups.

Significativité des variables :

Sexe : Le sexe a un effet significatif sur le nombre de sit-ups. Les hommes font en moyenne 2.79 sit-ups de moins que les femmes après ajustement pour les autres variables.

Âge : L'âge a un effet négatif très significatif sur les sit-ups, avec une diminution de 0.04 sit-ups par an, ce qui indique une diminution de l'endurance musculaire avec l'âge.

BMI : Le BMI n'a pas d'effet significatif (p-value = 0.260).

Pourcentage de graisse corporelle : Le pourcentage de graisse corporelle est marginalement significatif (p-value = 0.087) avec un effet négatif de -0.05 sit-ups, suggérant une légère réduction de la performance en sit-ups avec une plus grande masse grasse.

Conclusion pour Sit-ups :

Le nombre de sit-ups est fortement influencé par le sexe (les femmes effectuent plus de sit-ups) et par l'âge (les performances diminuent avec l'âge). Le BMI et le pourcentage de graisse corporelle n'ont qu'un faible effet sur cette variable.

4. Régression pour Flexibilité (Sit-and-Reach)

R² ajusté : 0.02683 (2.68% de la variance expliquée) Ce modèle explique une très faible proportion de la variance dans la flexibilité, ce qui montre que d'autres facteurs sont probablement en jeu.

SEE : 5.72 (Erreur standard d'estimation) L'erreur standard est de 5.72 cm, ce qui montre une erreur notable dans la prédiction des valeurs de flexibilité.

Significativité des variables :

Sexe : Le sexe a un effet significatif sur la flexibilité. Les hommes ont une flexibilité en moyenne 3.32 cm inférieure à celle des femmes.

Âge : L'âge a un effet négatif sur la flexibilité, avec une diminution de 0.02 cm par an.

BMI : Le BMI a un effet positif significatif sur la flexibilité, avec une augmentation de 0.13 cm pour chaque unité d'augmentation du BMI.

Pourcentage de graisse corporelle : Le pourcentage de graisse corporelle a également un effet significatif sur la flexibilité avec une diminution de 0.10 cm par unité d'augmentation.

Conclusion pour Flexibilité :

La flexibilité est influencée par le sexe (les femmes sont plus flexibles) et par le BMI (les personnes avec un BMI plus élevé ont une flexibilité légèrement supérieure). L'âge et le pourcentage de graisse corporelle affectent aussi la flexibilité de manière négative.

Phase 5 : Conclusion globale

Les modèles de régression montrent que le **sexe** est un facteur important pour la majorité des performances physiques mesurées (force de préhension, flexibilité, et endurance musculaire), et que l'**âge** joue un rôle majeur dans la diminution des capacités physiques (VO_2) et sit-ups). Le **BMI** n'a pas d'effet direct sur toutes les variables, mais il influence la flexibilité et la force de préhension, tandis que le **pourcentage de graisse corporelle** a un impact sur la capacité cardiovasculaire et la flexibilité.

Ces résultats suggèrent que les interventions pour améliorer la condition physique devraient prioriser l'entraînement musculaire et l'endurance chez les populations plus âgées, et que les différences entre sexes doivent être prises en compte dans les programmes d'entraînement.