

# Jeu de données *ToothGrowth* – Description et guide d’analyse

Travail de groupe – AIMS 2025/26

## Titre

L’effet de la vitamine C sur la croissance des dents chez les cobayes

## A. Description (traduction)

La réponse est la longueur des odontoblastes (cellules responsables de la croissance des dents) mesurée chez **60 cobayes**. Chaque animal a reçu une des trois doses de vitamine C (0,5, 1 et 2 mg/jour) administrées par l’une des deux méthodes suivantes : jus d’orange (0J) ou acide ascorbique (vitamine C pure, codé VC).

## Format

Un *data frame* de 60 observations sur 3 variables :

- `len` (numérique) : longueur de la dent (mm) ;
- `supp` (facteur) : type de supplément (VC ou 0J) ;
- `dose` (numérique) : dose en mg/jour (0,5, 1, 2).

## B. Explication et objectifs possibles

Le jeu de données sert à étudier l’effet de la dose et du type de supplément sur la croissance dentaire. Questions typiques :

- La longueur moyenne des dents diffère-t-elle entre les doses ? (effet dose)
- Le type de supplément (VC vs 0J) a-t-il un effet sur la longueur ? (effet supp)
- Existe-t-il une interaction dose × type de supplément ?

## C. Analyse recommandée (structure pour la présentation)

### 1. Analyse exploratoire (EDA)

- Résumés numériques : `summary(len)` ; moyennes, médianes, écart-type par groupe (`aggregate` ou `tapply`).
- Graphiques : histogrammes de `len`, boxplots de `len` selon `dose` et selon `supp`, scatter/stripchart si nécessaire.
- Tableaux : tableau croisé des effectifs par dose et supp.

## 2. Inférence

- Comparaisons de moyennes : tests t (paires) ou ANOVA à un facteur (dose) ; ANOVA à deux facteurs (dose et supp) pour tester interaction.
- Intervalles de confiance pour moyennes de chaque groupe.
- Régression linéaire (modèle avec `len` comme réponse, `dose` et `supp` comme facteurs) pour estimer effets et interactions.

## D. Exemple de code R (à utiliser pour l'analyse)

```
%           # Charger les données (dans R les données  
ToothGrowth sont fournies)  
data("ToothGrowth")  
str(ToothGrowth)  
summary(ToothGrowth)  
  
%           # R sum s par groupe  
aggregate(len ~ supp + dose, data = ToothGrowth,  
          FUN = function(x) c(mean=mean(x), sd=sd(x), n=  
length(x)))  
  
# Graphiques  
boxplot(len ~ dose, data = ToothGrowth, main="Len  
         selon la dose", xlab="Dose (mg/jour)", ylab="Longueur (mm)")  
boxplot(len ~ supp, data = ToothGrowth, main="Len  
         selon le suppl ment", xlab="Suppl ment",  
         ylab="Longueur (mm)")  
% Boxplot combin dose x supp  
interaction.plot(ToothGrowth$dose, ToothGrowth$  
                  supp, ToothGrowth$len, type="b", main="Interaction dose x supp", xlab="Dose", ylab="Longueur moyenne")  
  
%           # Test d'ANOVA (mod le deux facteurs avec  
interaction)  
ToothGrowth$dose <- factor(ToothGrowth$dose) #  
                   traiter dose comme facteur si souhait  
mod <- aov(len ~ supp * dose, data = ToothGrowth)  
summary(mod)  
  
# Si dose traite comme variable num rique (effet lin aire possible)  
mod2 <- lm(len ~ as.numeric(as.character(dose)) *  
           supp, data = ToothGrowth)  
summary(mod2)  
  
#Tests t : comparaison entre supp      dose = 0.5 par  
exemple  
subset05 <- subset(ToothGrowth, dose == 0.5)
```

```

t.test(len ~ supp, data = subset05)

# Intervalles de confiance pour moyennes (ex.
# pour OJ)
library(dplyr)
ToothGrowth %>% filter(supp == "OJ") %>%
  summarise(mean_len = mean(len), sd = sd(len), n =
  n(),
  se = sd/sqrt(n),
  ci_low = mean_len - qt(0.975, n-1)*se,
  ci_high = mean_len + qt(0.975, n-1)*se)

```

## E. Exemples d'interprétation (à présenter)

- Si l'ANOVA montre un effet significatif de la dose : « la longueur moyenne augmente significativement avec la dose. »
- Si l'effet de `supp` est significatif : « le type de supplément (OJ vs VC) influence la croissance. »
- Si l'interaction est significative : « l'effet de la dose dépend du type de supplément (par ex. OJ produit une plus grande augmentation à certaines doses). »

## F. Conseils pratiques pour la présentation (12 minutes)

1. 0–2 min : Introduction et description du jeu de données.
2. 2–6 min : EDA — graphiques clés et résumés numériques (1–2 figures).
3. 6–11 min : Inférence — poser la question principale, méthode choisie, résultats (tableau ou sortie ANOVA / test t), interprétation.
4. 11–12 min : Conclusion concise et recommandations (limites de l'étude, perspectives).

**Bonne analyse !** Si tu veux, je peux :

- te fournir le code R complet exécuté avec sorties et graphiques (si tu veux un PDF des graphiques),
- ou convertir ce document en PDF/Beamer pour la présentation.

## Instructions

- À chaque groupe a été attribué un jeu de données différent. Tous les jeux de données comprennent :
  - au moins une **variable catégorielle** qui divise les données en deux groupes ou plus ;
  - quelques **variables continues**.
- Les détails sur votre jeu de données spécifique sont fournis dans la feuille `DataDescription` du tableauur qui vous a été attribué.
- Votre tâche pour la présentation est d'**analyser les données**. Des orientations sur les questions à traiter sont fournies ci-dessous.

- La présentation doit être divisée en deux parties :
    1. **Analyse exploratoire des données (EDA)** : statistiques descriptives du jeu de données.
    2. **Inférence** : questions statistiques de votre choix.
- Veillez à relier correctement ces deux parties dans votre présentation.

## Directives générales

Voici quelques lignes directrices sur le type de questions que vous pouvez explorer. Adaptez ou étendez selon votre jeu de données.

### 1. Analyse exploratoire des données (EDA)

- Utilisez à la fois des **résumés numériques** et des **résumés graphiques** pour décrire vos données.
- Lorsque vous présentez des résumés numériques, réfléchissez si un **tableau** ne les rendrait pas plus clairs.
- Si vous créez plusieurs graphiques pour la même variable, mettez-les en relation entre eux et avec les résultats numériques.
- Personnalisez vos graphiques (étiquettes d'axes, taille des polices, couleurs, légendes, etc.). Les graphiques doivent être clairs et informatifs pour l'auditoire.

### 2. Inférence

- Dans votre présentation, **énoncez clairement** les questions que vous voulez répondre et justifiez les méthodes choisies.
- Vous pouvez construire des **intervalles de confiance** pour certains paramètres et comparer les résultats entre groupes si pertinent.
- Les intervalles de confiance s'appliquent aussi aux **proportions** — vérifiez la pertinence pour votre jeu de données.
- Effectuez des **tests d'hypothèses** pour comparer des groupes lorsque c'est approprié.
- Si pertinent, explorez les **relations entre variables numériques** (corrélation, régression linéaire, ...).
- Indiquez les formules utilisées et présentez les résultats ; préparez-vous à les commenter et à expliquer leur sens.

**Remarque importante :** Les points ci-dessus sont des suggestions. Les questions les plus pertinentes dépendront de votre jeu de données. N'en faites pas trop : vous disposez de **12 minutes** pour la présentation — privilégiez la clarté.

## Considérations supplémentaires

- Pendant l'EDA, vérifiez si l'une des distributions étudiées convient visuellement à vos données (superposer une densité théorique sur un histogramme).

- Si une distribution semble convenir, complétez le contrôle visuel par un contrôle numérique : comparez quelques percentiles observés avec les percentiles théoriques correspondants.

## Note finale

Privilégiez la clarté et l'interprétation plutôt que la quantité. Une analyse concise et bien raisonnée est plus efficace qu'une présentation surchargée.