

R, premiers pas

Tester une hypothèse

David Causeur
Institut Agro Rennes Angers
IRMAR UMR 6625 CNRS

26 janvier, 2025

Objectifs de la formation

Deux types d'objectifs :

- ▶ Acquérir des compétences d'analyse de données
 - ▶ Méthodes pour décrire et démontrer l'effet d'une variable sur une autre
 - ▶ Démarches pour conduire une analyse complexe, construire un modèle intégrant plusieurs effets

- ▶ Se familiariser avec R et Rstudio
 - ▶ Mettre en oeuvre
 - ▶ Editer des rapports

Modalités pédagogiques

- ▶ Apprentissage par mise en situation
- ▶ Formalisation mathématique réduite au minimum
- ▶ Programme modulable selon vos questions

Environnements R et Rstudio

 : environnement pour l'analyse de données et la visualisation

- ▶ R est un logiciel libre
- ▶ Site web : <https://www.r-project.org>
- ▶ Jan. 2025 : 22 000 packages <https://cran.r-project.org/>
- ▶ Communautés R :
 - ▶ <https://www.r-bloggers.com/>,
 - ▶ <https://rladies.org/>,
 - ▶ <https://r-toulouse.netlify.app/>
 - ▶ <https://stackoverflow.com/questions/>
 - ▶ <https://r-graph-gallery.com/>



: interface pour R (IDE : Integrated Development Environment)

- ▶ Site web : <https://posit.co/>
- ▶ Environnement ergonomique pour l'analyse de données
- ▶ Outils pour l'édition de rapport d'études
 - ▶ Rmarkdown : <http://rmarkdown.rstudio.com/>
- ▶ Préparation de la session de travail :
 - ▶ Créez un projet pour l'analyse des données **fruit**
 - ▶ Installez (si besoin) le package **tidyverse**

Importer des données

```
fruit <- read.table(file="data/fruit.txt",stringsAsFactors=TRUE)
str(fruit)
```

```
'data.frame':  435 obs. of  12 variables:
 $ Poids      : num  48 40.2 42 47.4 35.2 60.5 66.7 44.1 44.5 43.6 ...
 $ Diam       : num  36.9 36.9 35.4 37.5 33.9 ...
 $ L          : num  64.2 58.1 61 56.4 59.7 ...
 $ a          : num  -6.78 -13.2 -6.71 5.43 -12.31 ...
 $ b          : num  42.4 39.1 36.8 34.8 40.3 ...
 $ Glucose    : num  1.28 1.3 2.69 2.28 1.17 1.4 1.19 1.34 1.24 1.21 ...
 $ Fructose   : num  0.52 0.47 0.91 0.74 0.5 0.57 0.46 0.51 0.52 0.58 ...
 $ Saccharose: num  2.48 1.71 5.46 4.27 1.72 4.01 3.33 2.29 2.82 1.74 ...
 $ Citrate    : num  36.8 37.9 38.5 34.9 36.1 ...
 $ Malate     : num  3.95 4.98 4.57 5.33 4.15 4.56 4.94 5.24 4.6 4.04 ...
 $ Variete    : Factor w/ 4 levels "37","bl","go",...: 1 1 1 1 1 1 1 1 1 1 ...
 $ Maturite   : Factor w/ 3 levels "Faible","Forte",...: 1 1 1 1 1 1 1 1 1 1 ...
```

Tester un effet

Principes généraux

Y-a t'il un effet de *ceci sur cela* ? : lien entre ceci et cela à l'échelle d'une population

Illustration :

- ▶ Les poids moyens sont-ils les mêmes pour toutes les variétés ?
- ▶ Le diamètre dépend-il de l'indice de clarté L ?

Dans ces deux problématiques, que sont :

- ▶ la population visée ?
- ▶ ceci (variable explicative) ?
- ▶ cela (variable à expliquer) ?

Effet groupe

Les poids moyens des fruits par variété sont-ils différents ?

```
tab <- fruit %>% group_by(Variete) %>%  
  summarise(Mean=mean(Poids),Sd=sd(Poids),  
            Min=min(Poids),Max=max(Poids))  
tab %>% kbl(caption="Statistiques élémentaires pour le poids") %>%  
  kable_paper(full_width = F)
```

Statistiques élémentaires pour le poids

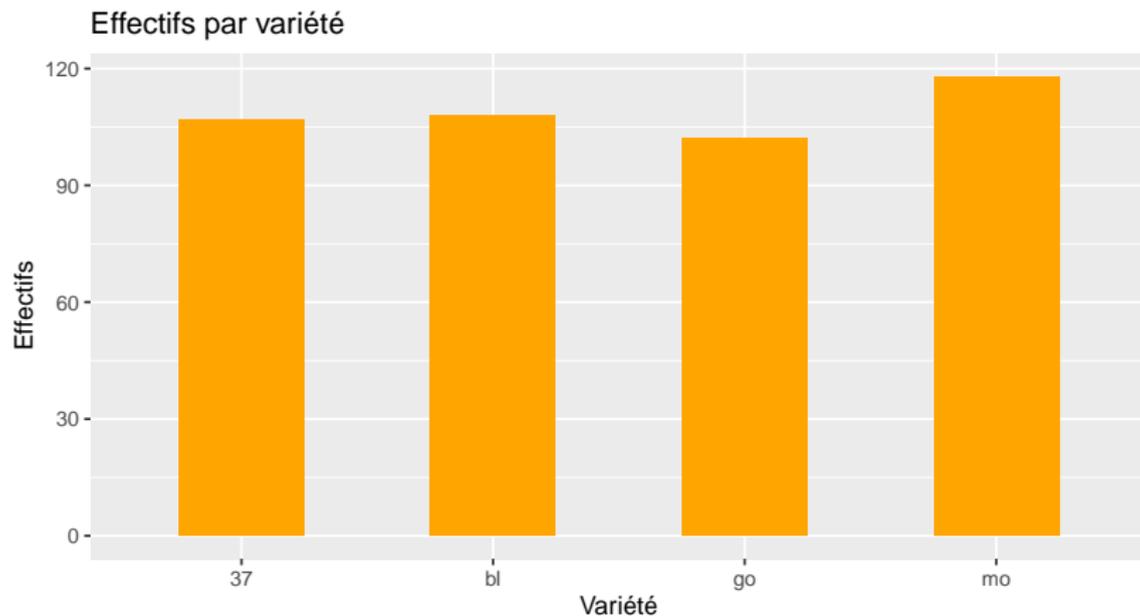
Variete	Mean	Sd	Min	Max
37	55.98598	9.428074	35.2	73.5
bl	53.16019	7.764741	38.7	77.4
go	61.62059	10.625696	41.3	96.7
mo	70.66356	9.908070	47.3	103.9

Les poids moyens par variété **sont différents**.

Le sont-ils **suffisamment** pour conclure que les différences entre variétés existent à l'échelle de toute la production ?

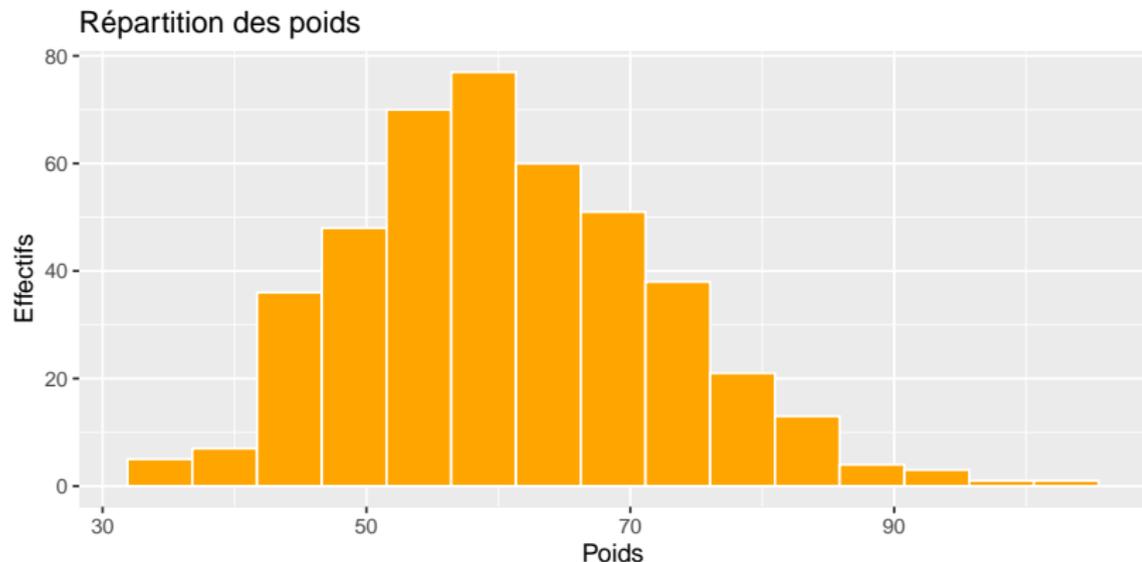
Représentation des variétés dans l'échantillon

```
fruit %>% ggplot() +  
  geom_bar(aes(x = Variete),fill="orange",width=0.5) +  
  ggtitle("Effectifs par variété") + xlab("Variété") + ylab("Effectifs")
```



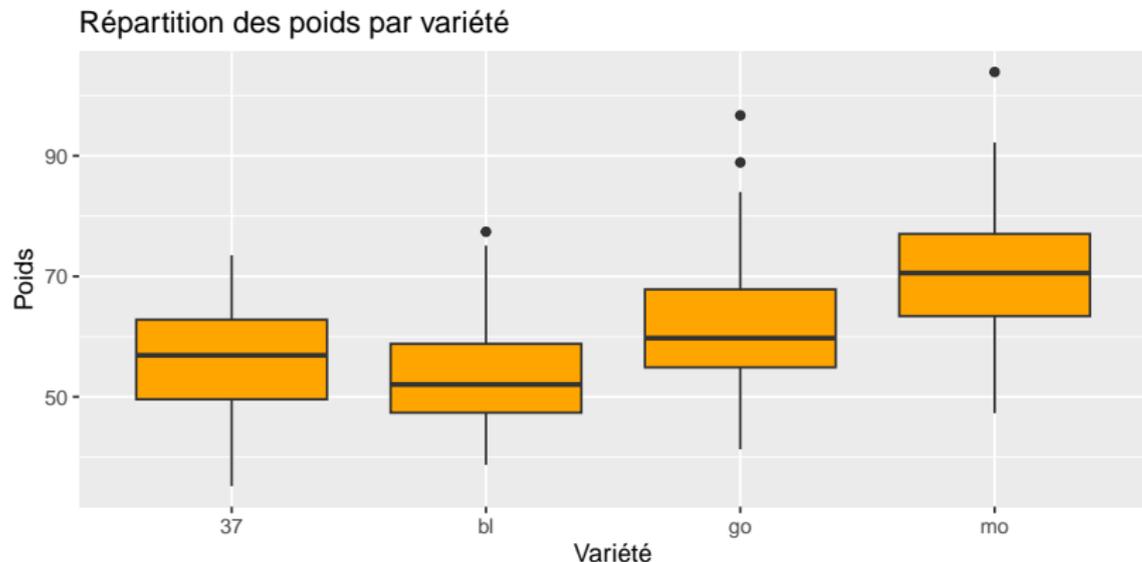
Répartition des poids des fruits dans l'échantillon

```
fruit %>% ggplot() + aes(x = Poids) +  
  geom_histogram(fill="orange",bins=15,col="white") +  
  ggtitle("Répartition des poids") + xlab("Poids") +  
  ylab("Effectifs")
```



Répartition des poids par variété

```
fruit %>% ggplot() +  
  geom_boxplot(aes(x=Variete,y=Poids),fill="orange") +  
  ggtitle("Répartition des poids par variété") +  
  xlab("Variété") + ylab("Poids")
```



Modèle d'analyse de la variance à un facteur

Ajustement du modèle :

```
mod1 <- lm(Poids~Variete,data=fruit)
coef(mod1)
```

(Intercept)	Varietebl	Varietego	Varietemo
55.985981	-2.825796	5.634607	14.677578

Quelle est la variété pour laquelle les fruits sont les plus lourds ?

Sous l'hypothèse d'absence effet (modèle nul)

```
mod0 <- lm(Poids~1,data=fruit)
coef(mod0)
```

(Intercept)
60.58713

Pourquoi l'estimation de '(Intercept)' n'est pas la même avec les deux modèles ?

Comparaison des modèles

```
anova(mod0,mod1)
```

Analysis of Variance Table

Model 1: Poids ~ 1

Model 2: Poids ~ Variete

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	434	59075				
2	431	38763	3	20313	75.284	< 2.2e-16

Conclusion : les poids moyens par variété sont significativement différents au seuil de 5%,



! : Les poids moyens par variété ne sont peut-être pas **tous** mutuellement significativement différents.

Comparaison de deux moyennes

Les poids moyens des fruits de variétés A et B sont-ils significativement différents ?

- ▶ **Test de Student** : pour des hypothèses nulles du type **un paramètre est égal à une valeur cible**
- ▶ Ici, l'hypothèse nulle est : **la différence entre les poids moyens des fruits de variétés A et B est égale à 0**

```
fruit2 <- fruit %>% filter(Variete%in%c("37","go")) %>% droplevels()
t.test(Poids~Variete,data=fruit2,var.equal=TRUE)
```

Two Sample t-test

data: Poids by Variete

t = -4.0595, df = 207, p-value = 6.973e-05

alternative hypothesis: true difference in means between group 37 and group go is not equal

95 percent confidence interval:

-8.371071 -2.898143

sample estimates:

mean in group 37	mean in group go
55.98598	61.62059

Puissance du dispositif experimental

Illustration : Si les poids moyens des fruits de variétés A et B diffèrent de $\delta = 2$ grammes, le test détecte-t'il cette différence ?

```
mod = lm(Poids~Variete,data=fruit2)
sigma = summary(mod)$sigma
power.t.test(delta=2,n=100,sd=sigma,sig.level=0.05)$power
```

```
[1] 0.2888122
```

Combien de fruits sont nécessaires pour qu'une différence de 2g soit détectable ?

Comparaisons par paires

Illustration : quelles sont les variétés dont les poids moyens sont significativement différents ?

```
comp <- emmeans(mod1, ~ Variete)
pairs(comp, adjust="bonf")
```

contrast	estimate	SE	df	t.ratio	p.value
37 - bl	2.83	1.29	431	2.185	0.1768
37 - go	-5.63	1.31	431	-4.294	0.0001
37 - mo	-14.68	1.27	431	-11.594	<.0001
bl - go	-8.46	1.31	431	-6.461	<.0001
bl - mo	-17.50	1.26	431	-13.860	<.0001
go - mo	-9.04	1.28	431	-7.053	<.0001

P value adjustment: bonferroni method for 6 tests

Comparaison graphique des variétés de fruits selon leur poids

```
res <- meansComp(mod1, ~ Variete, graph=TRUE)
```

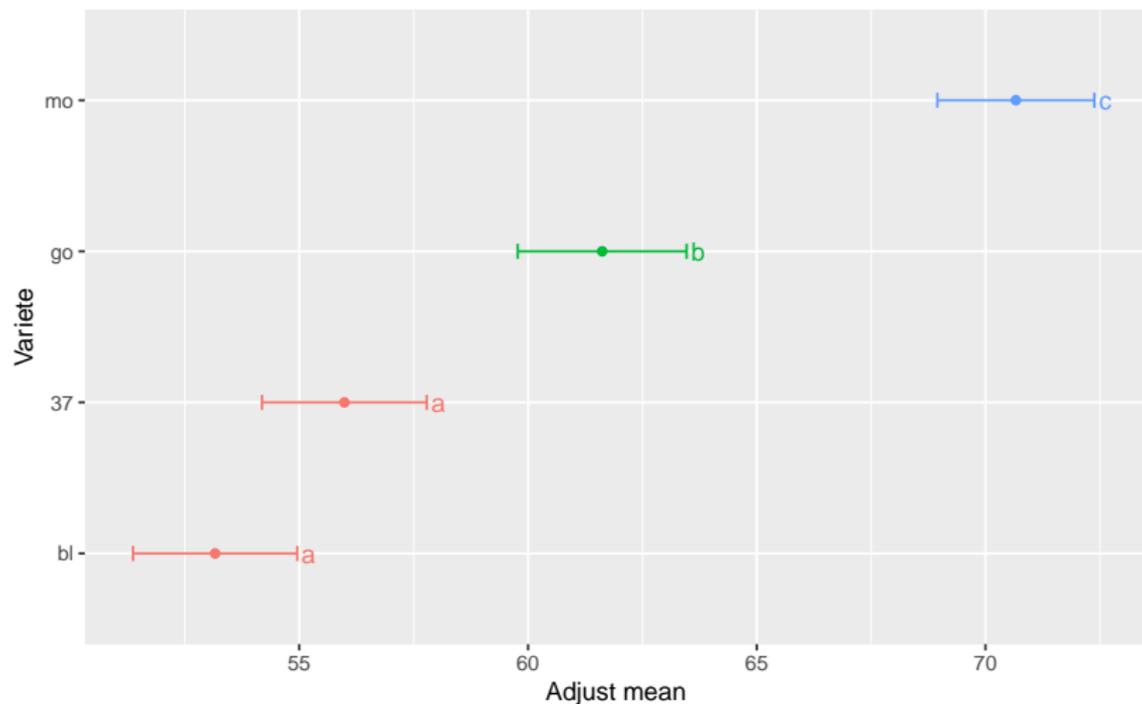
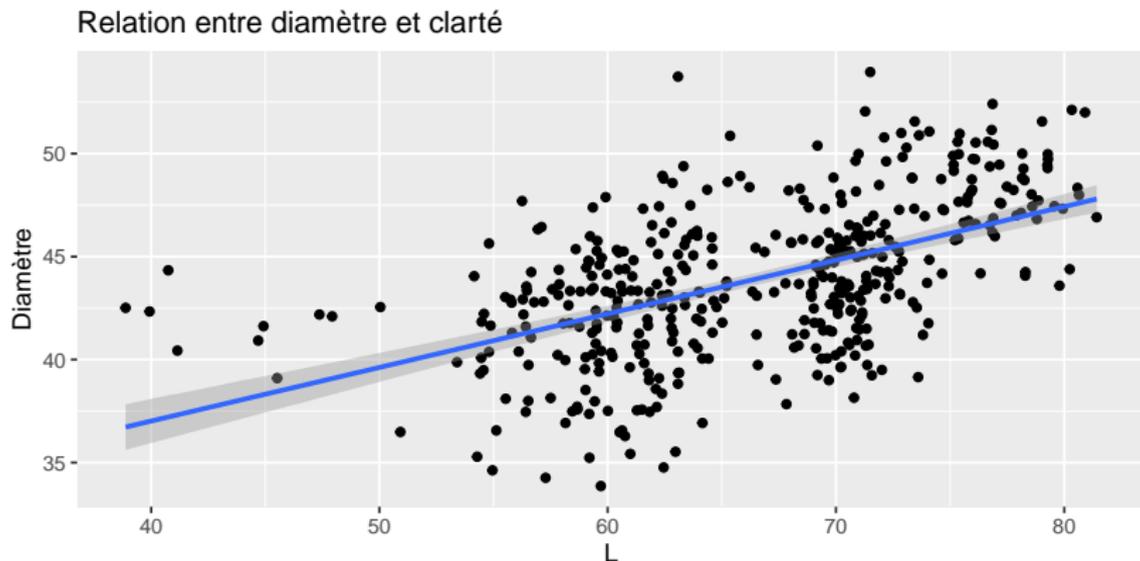


Illustration : les indices 'a' moyens par stade de maturité sont-ils différents ?

Effet linéaire

Diamètre d'un fruit en fonction de son indice de clarté

```
fruit %>% ggplot() + aes(x=L,y=Diam) +  
  geom_point() + geom_smooth(method='lm') +  
  ggtitle("Relation entre diamètre et clarté") +  
  xlab("L") + ylab("Diamètre")
```



Coefficient de corrélation : $-1 \leq r \leq 1$:

```
cor(fruit$L,fruit$Diam)
```

```
[1] 0.5350462
```

Interprétation :

- ▶ Si $r \approx 1$: relation linéaire croissante très forte
- ▶ Si $r \approx -1$: relation linéaire décroissante très forte
- ▶ Si $r \approx 0$: absence de lien linéaire

Quelle est la variable la plus corrélée au taux de saccharose parmi le poids, le diamètre, L, a et b ?

Modèle de régression linéaire

Estimation de la droite de régression

```
mod1 <- lm(Diam~L,data=fruit)
coef(mod1)
```

```
(Intercept)          L
 26.6062360    0.2602886
```

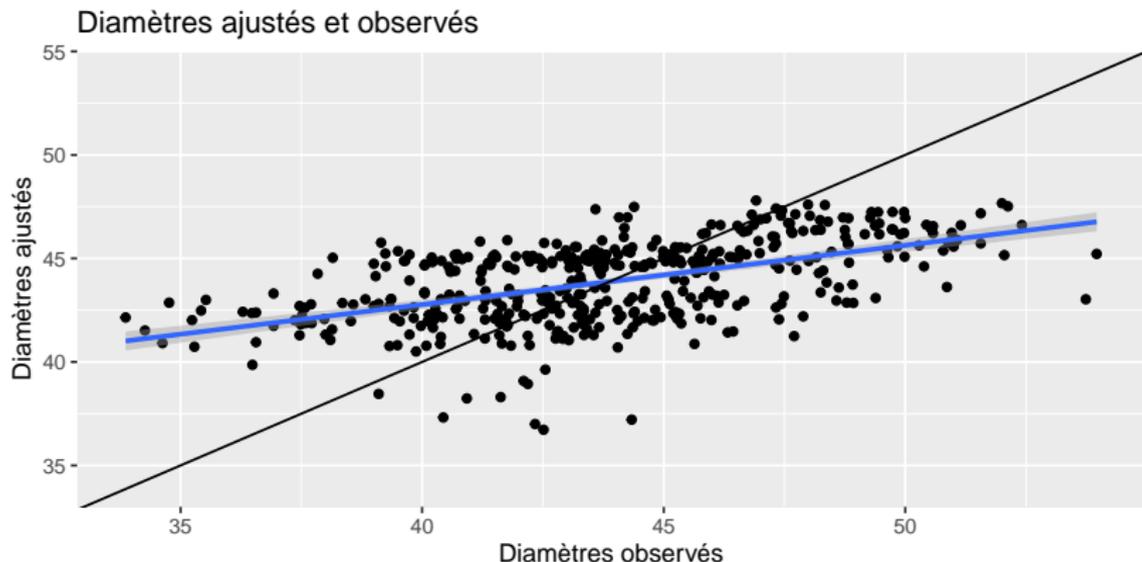
Evaluation numérique de la qualité de l'ajustement : R^2

```
summary(mod1)$r.squared
```

```
[1] 0.2862744
```

Evaluation graphique de la qualité de l'ajustement

```
fruit$Diam_fit <- fitted(mod1)
fruit %>% ggplot(aes(x=Diam,y=Diam_fit)) + geom_point() + geom_smooth(method='lm') +
  ggtitle("Diamètres ajustés et observés") + geom_abline(intercept=0,slope=1) +
  xlab("Diamètres observés") + ylab("Diamètres ajustés") + ylim(34,54)
```



Comparaison des modèles avec et sans effet

```
mod0 <- lm(Diam~1,data=fruit)
anova(mod0,mod1)
```

Analysis of Variance Table

Model 1: Diam ~ 1

Model 2: Diam ~ L

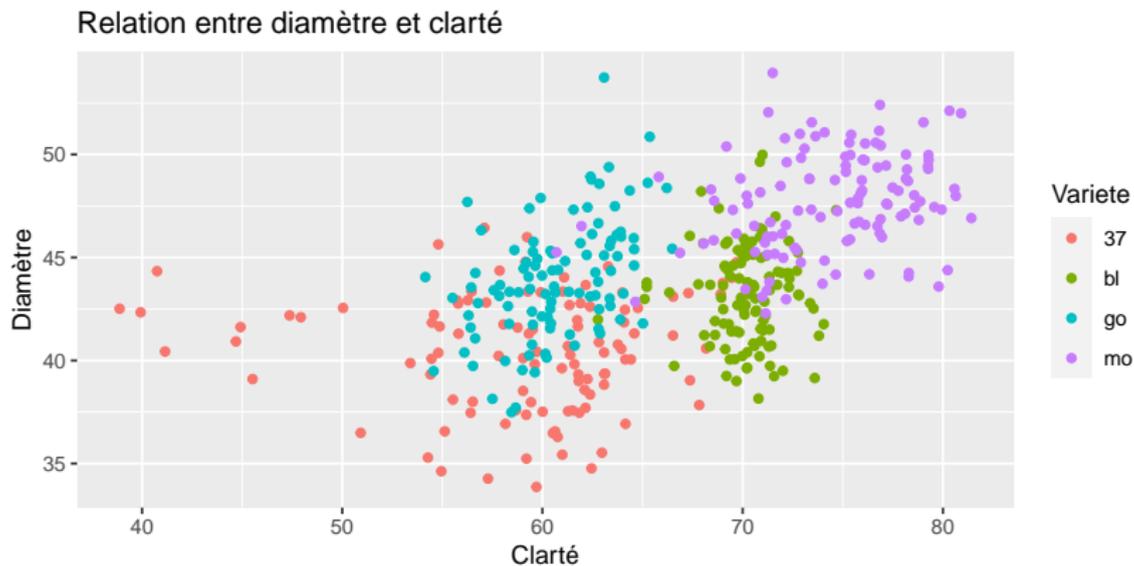
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	434	5987.5				
2	433	4273.5	1	1714.1	173.68	< 2.2e-16

Conclusion : la relation linéaire entre diamètre et clarté est significative au seuil de 5%,

Effet linéaire par groupes

Visualisation d'un effet linéaire par groupes

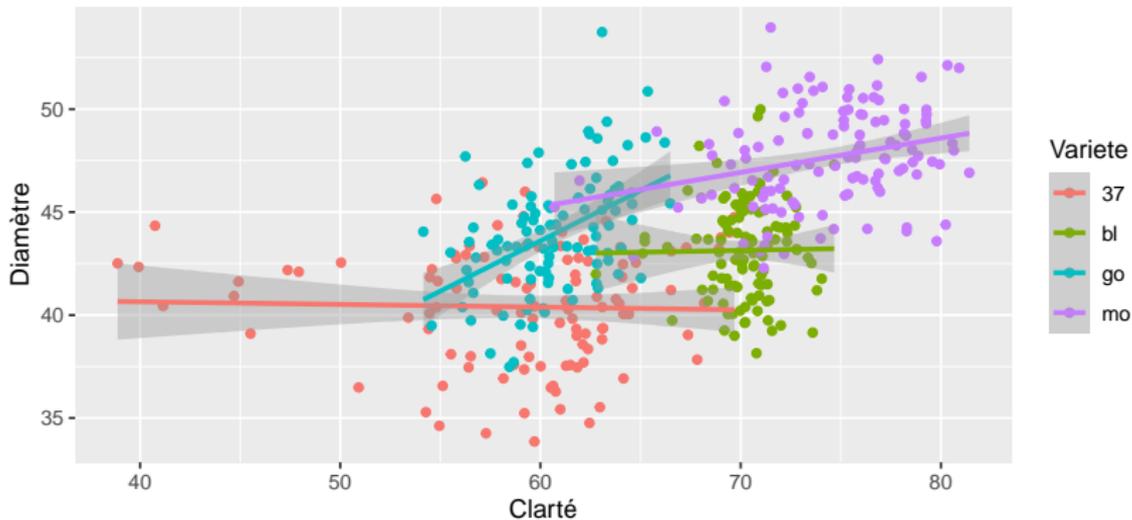
```
p <- fruit %>% ggplot() + geom_point(aes(x=L,y=Diam,col=Variete)) +  
  ggtitle("Relation entre diamètre et clarté") + xlab("Clarté") + ylab("Diamètre")  
p
```



Tendance linéaire par groupes

```
p + geom_smooth(method="lm", aes(x=L, y=Diam, col=Variete))
```

Relation entre diamètre et clarté



Modèle pour un effet linéaire par groupes

```
mod2 <- lm(Diam~L*Variete,data=fruit)
summary(mod2)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	41.16861452	2.33378621	17.64026814	1.038707e-52
L	-0.01301521	0.03930626	-0.33112316	7.407139e-01
Varieteb1	0.72726879	9.41825478	0.07721906	9.384854e-01
Varietego	-26.86983304	6.05374842	-4.43854471	1.154377e-05
Varietemo	-5.88749675	4.96358662	-1.18613761	2.362275e-01
L:Varieteb1	0.03078736	0.13572209	0.22684122	8.206558e-01
L:Varietego	0.50147283	0.09998262	5.01559995	7.767680e-07
L:Varietemo	0.17937748	0.07072915	2.53611810	1.156430e-02

Quelle est l'équation permettant de prédire le diamètre par la clarté pour la variété A ?
et pour la variété B ?

Equation unique ou équations par groupe ?

Test de Fisher

```
anova(mod1,mod2)
```

Analysis of Variance Table

Model 1: Diam ~ L

Model 2: Diam ~ L * Variete

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	433	4273.5				
2	427	2717.4	6	1556	40.751	< 2.2e-16

- ▶ Equation globale ou équations par variété ?
- ▶ La relation entre le poids et le diamètre est-elle la même pour toutes les variétés ?

Choix du meilleur modèle

Prédiction du taux de sucre d'un fruit

Modèle de régression linéaire multiple

```
mod <- lm(Saccharose~Poids+Diam+L+a+b,data=fruit)
coef(mod)
```

(Intercept)	Poids	Diam	L	a	b
-8.78354474	0.02604814	-0.01050177	0.17831239	0.09508650	-0.01578916

Qualité d'ajustement

```
summary(mod)$r.squared
```

```
[1] 0.4638385
```

Test de Fisher

```
mod0 = lm(Saccharose~1,data=fruit)
anova(mod0,mod)
```

Analysis of Variance Table

Model 1: Saccharose ~ 1

Model 2: Saccharose ~ Poids + Diam + L + a + b

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	434	1140.9				
2	429	611.7	5	529.19	74.226	< 2.2e-16

Quelle conclusion en tirer ?

Confusion d'effets

Quels sont les effets à retenir ?

```
car::Anova(mod)
```

Anova Table (Type II tests)

Response: Saccharose

	Sum Sq	Df	F value	Pr(>F)
Poids	7.02	1	4.9265	0.02697
Diam	0.09	1	0.0648	0.79920
L	274.77	1	192.7026	< 2e-16
a	223.81	1	156.9652	< 2e-16
b	3.95	1	2.7703	0.09676
Residuals	611.70	429		

Le diamètre apporte-t'il une information sur le taux de sucre ?

Choix du meilleur modele

Quel choix de variables explicatives ?

```
bestmod <- regsubsets(Saccharose-Poids+Diam+L+a+b,data=fruit)
tidy(bestmod) %>% dplyr::select(1:7) %>%
  kbl(caption="Recherche exhaustive du meilleur modèle") %>%
  kable_paper(full_width = F)
```

Recherche exhaustive du meilleur modèle

(Intercept)	Poids	Diam	L	a	b	r.squared
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	0.1907737
TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	0.4328652
TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	0.4602912
TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	0.4637576
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	0.4638385

Compromis entre qualité d'ajustement et complexité du modèle

Critère d'information d'un modèle : BIC (Bayesian Information Criterion) et AIC (Akaike IC)

- ▶ Mesure de la qualité d'ajustement d'un modèle tenant compte de sa complexité (nombre de paramètres)

```
tidy(bestmod) %>% dplyr::select(c(1:6,9)) %>%  
  kbl(caption="Recherche exhaustive du meilleur modèle") %>%  
  kable_paper(full_width = F)
```

Recherche exhaustive du meilleur modèle

(Intercept)	Poids	Diam	L	a	b	BIC
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	-79.92869
TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	-228.48783
TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	-243.97423
TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	-240.70175
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	-234.69210

Sélection pas à pas

```
mod <- lm(Saccharose~Poids+Diam+L+a+b,data=fruit)
select <- stepwise(mod,direction="forward/backward",
                  criterion="BIC",trace=0)
```

Direction: forward/backward

Criterion: BIC

```
coef(select)
```

(Intercept)	Diam	L	a
-10.82598543	0.07574545	0.16615627	0.09227773

Le sens de la recherche affecte-t'il le résultat ?

Ce qu'il faut retenir

L'identification d'une relation entre une variable réponse et une ou plusieurs variables explicatives (un effet) repose sur les principes suivants :

- ▶ L'effet d'intérêt peut être visualisé par le choix adapté de graphiques, impliquant les variables directement impliquées mais aussi des variables pouvant porter des effets de confusion ;
- ▶ De même, les tests d'hypothèse doivent être mis en oeuvre dans des modèles ajustant des effets de confusion ;
- ▶ La sélection de tous les effets à retenir pour décrire les variations de la variable réponse est possible par des procédures visant à trouver le meilleur compromis entre qualité d'ajustement et complexité du modèle.

Pour aller plus loin

- ▶ Les principes généraux donnés dans le cadre du modèle linéaire restent valides pour des modèles plus complexes
 - ▶ Voir modélisation non-linéaire, modèle linéaire généralisé