

Étude Statistique : Relations entre Tabagisme et problème pulmonaire

BOUFAGHER Mohamed
HAMADACHE Youcef
MECHTOUH Nacer
AMARATUNGA Mélanie

Introduction :

Les maladies respiratoires chroniques, et en particulier les troubles pulmonaires, représentent aujourd’hui un enjeu majeur de santé publique. Dans de nombreux pays industrialisés, leur prévalence n’a cessé d’augmenter au cours des dernières décennies, notamment en raison de l’exposition croissante aux facteurs environnementaux et comportementaux à risque. L’impact de ces maladies est considérable, tant sur le plan individuel que collectif, avec des coûts économiques et sociaux importants pour les systèmes de santé.

Parmi les facteurs de risque les plus fréquemment identifiés dans la littérature médicale, le tabagisme actif, mais aussi le tabagisme passif, occupent une place centrale. Leurs effets néfastes sur les voies respiratoires sont aujourd’hui bien documentés. Cependant, ces comportements ne suffisent pas à eux seuls à expliquer l’ensemble des cas observés. Des caractéristiques individuelles telles que l’âge, le sexe ou encore la situation conjugale pourraient également influencer la vulnérabilité d’un individu face aux problèmes pulmonaires, soit en renforçant, soit en atténuant l’impact des expositions toxiques.

Dans ce contexte, la présente étude propose une analyse statistique approfondie visant à identifier les variables individuelles susceptibles d’augmenter la probabilité de développer des problèmes pulmonaires. L’approche méthodologique adoptée repose sur l’exploitation d’un jeu de données composé de 100 individus, chacun décrit par des variables démographiques (âge, sexe), sociales (situation familiale) et comportementales (consommation de tabac, exposition au tabagisme passif). La variable cible de l’analyse est binaire, indiquant la présence ou l’absence de symptômes respiratoires diagnostiqués ou suspectés.

Pour répondre à cette problématique, une régression logistique a été mise en œuvre. Cette technique permet d’estimer la probabilité qu’un individu souffre de problèmes pulmonaires en fonction de ses caractéristiques personnelles. En complément, des analyses descriptives et graphiques ont été réalisées afin d’explorer les distributions de chaque variable, ainsi que les relations apparentes entre celles-ci et l’état pulmonaire des individus. Les résultats obtenus permettront non seulement de confirmer certaines hypothèses médicales déjà bien établies, mais aussi d’apporter un éclairage supplémentaire sur les facteurs qui méritent une attention particulière dans la mise en place de politiques de prévention.

1. Préparation et compréhension du jeu de données.

L'analyse commence par l'importation du jeu de données à l'aide de la commande `read.csv("2_tabac.csv", stringsAsFactors = FALSE)` qui permet d'éviter que les chaînes de caractères ne soient transformées automatiquement en facteurs. Cela facilite une manipulation plus flexible des variables par la suite. L'objet `data_original` contient 100 observations sur 9 variables

Pour examiner la structure du jeu de données, on utilise `str(data_original)` ce qui révèle que la variable `probleme_pulmonaire` est de type logique, prenant les valeurs TRUE ou FALSE. Cette structure est appropriée pour une analyse binaire, mais elle devra être recodée pour certaines fonctions. La commande permet aussi de visualiser les premiers niveaux des autres variables comme `sexe`, `tabac` ou `situation` et de vérifier rapidement qu'elles sont bien codées

La commande `colSums(is.na(data_original))` nous informe sur la présence de valeurs manquantes. Ici, elle retourne uniquement des zéros, indiquant que toutes les cellules sont complètes. Aucun nettoyage additionnel n'est requis à cette étape

Avec `summary(data_original)`, on obtient un aperçu global de chaque variable. Par exemple, pour `age`, on observe une médiane autour de 41 ans, avec un intervalle allant de 18 à 73 ans. Pour `probleme_pulmonaire`, le résumé montre que 35 individus présentent un problème (TRUE) contre 65 qui n'en présentent pas (FALSE). Cela correspond à 35 % de la population, ce qui suggère une prévalence relativement élevée dans l'échantillon. Voici la sortie total de `summary(data_original)` :

```
> summary(data_original)
   ID          age         sexe      situation      tabac
Min. : 0.00  Min. :20.00  femme:46  celibataire:39  0     :51
1st Qu.:24.75 1st Qu.:29.75  homme:54  en_couple  :36  1     :6
Median :49.50 Median :41.00        marié  :20  13    :6
Mean   :49.50 Mean  :41.38        veuf   : 5  6    :5
3rd Qu.:74.25 3rd Qu.:52.25                    9    :5
Max.   :99.00 Max.  :64.00                    11   :5
                                         (Other):22
tabagisme_passif probleme_pulmonaire classe_tabac      proba
FALSE:84          0:65           Non-fumeur: 0  Min.  :0.0000
TRUE :16          1:35           Faible    :64  1st Qu.:0.0000
                           Moyen    :18  Median :0.0248
                           Fort    :18  Mean   :0.3500
                                         3rd Qu.:0.9741
                                         Max.   :1.0000
```

Avant la modélisation, la variable cible est recodée à l'aide de `ifelse(data_original$probleme_pulmonaire == TRUE, 1, 0)`, ce qui la convertit en variable binaire numérique. Ensuite, elle est transformée en facteur binaire avec `factor(..., levels = c(0, 1))`, comme requis pour la régression logistique et la courbe ROC.

Les autres variables comme `sexe`, `situation` ou `tabagisme_passif` sont converties en facteurs via `as.factor()` afin de garantir leur traitement comme variables qualitatives. La variable

tabac, qui indique le niveau de consommation, est classée en quatre catégories à l'aide de la fonction **cut()**. Cela permet de différencier les non-fumeurs, les fumeurs faibles, moyens et forts. Ce regroupement rend les analyses visuelles plus claires et plus parlantes. Voici les lignes de code qui nous permettent cela :

```
# Conversion des variables explicatives en facteur si nécessaire
data_original$sexe <- as.factor(data_original$sexe)
data_original$situation <- as.factor(data_original$situation)
data_original$tabac <- as.factor(data_original$tabac)
data_original$tabagisme_passif <- as.factor(data_original$tabagisme_passif)

# Création d'une variable catégorisée pour le tabac
data_original$classe_tabac <- cut(as.numeric(data_original$tabac),
                                    breaks = c(-1, 0, 5, 10, 15),
                                    labels = c("Non-fumeur", "Faible", "Moyen", "Fort"))
```

(Nous avons appris à utiliser cette manipulation grâce à la fonction **cut()** qui permet de créer des tranches à partir d'une variable numérique, ici convertie avec **as.numeric(data_original\$tabac)**. Les bornes spécifiées via **breaks = c(-1, 0, 5, 10, 15)** segmentent les niveaux de consommation en quatre catégories distinctes. En associant les libellés à ces intervalles avec **labels = c("Non-fumeur", "Faible", "Moyen", "Fort")**, nous avons pu générer une variable interprétable qui facilite la modélisation et la visualisation. Cette méthode, que nous ne connaissions pas au départ, s'est révélée essentielle pour classer les individus selon leur profil tabagique et améliorer l'interprétation des résultats)

Ainsi, cette préparation permet d'obtenir un tableau de données propre, cohérent et utilisable pour des analyses plus avancées. C'est une étape fondamentale pour assurer la fiabilité des résultats futurs car une mauvaise structuration ici aurait pu fausser les modèles ou les interprétations

2. Visualisation graphiques & interprétation :

L'analyse visuelle constitue une étape essentielle dans la compréhension des relations entre les variables explicatives et la variable cible. Nous avons commencé par tracer un histogramme du niveau de tabac après conversion en numérique : ce graphique met en évidence une forte concentration de non-fumeurs dans l'échantillon alors que les fumeurs se répartissent entre niveaux faibles, moyens et forts. Cette distribution déséquilibrée nous a conduit à créer des classes afin de simplifier la lecture et d'améliorer l'interprétation des résultats.

Le boxplot de l'âge selon la présence de problèmes pulmonaires permet d'observer une tendance intéressante. Les individus atteints de troubles respiratoires ont un âge médian plus élevé que ceux qui en sont exempts : cela suggère que l'âge pourrait jouer un rôle dans l'apparition des problèmes pulmonaires, hypothèse qui sera testée dans la modélisation. (Voir annexe page 1 n°1)

Pour examiner le lien entre tabagisme et santé pulmonaire, nous avons construit un graphique en barres empilées, représentant la répartition des individus avec ou sans problèmes pulmonaires selon leur classe de tabac. Ce graphique est issu du code suivant : (Voir annexe page 1 n°2)

```
# Barplot classe de tabac vs problème
ggplot(data_original, aes(x = factor(probleme_pulmonaire), fill = classe_tabac)) +
  geom_bar(position = "fill") +
  labs(title = "Problèmes pulmonaires selon le niveau de tabagisme",
       x = "Problème pulmonaire", y = "Proportion", fill = "Classe de tabac") +
  scale_fill_brewer(palette = "Set2") +
  theme_minimal()
```

Nous avons choisi de regrouper les individus par classe de consommation de tabac (non-fumeur, faible, moyen, fort) car une première tentative sans ces regroupements produisait un graphique difficile à lire, chaque niveau de tabac étant représenté individuellement, ce qui surchargeait la visualisation et n'apportait pas de lisibilité supplémentaire. En classant les individus, nous avons obtenu un graphique plus synthétique, qui met en lumière une forte proportion de fumeurs moyens et forts chez les individus atteints de problèmes pulmonaires, tandis que les non-fumeurs sont majoritaires dans le groupe sans problème

Le choix d'utiliser les classes nous a donc permis de produire une représentation clair des relations entre consommation de tabac et santé respiratoire. Le graphique permet ainsi de visualiser une tendance nette : plus la consommation de tabac est élevée, plus la probabilité de développer des troubles respiratoires semble augmenter

Voici le code qui montre comment le graphique apparaissait initialement, sans regroupement par classe, et qui illustre pourquoi cette version était beaucoup moins lisible :

```
# Graphique : Illisible pour la répartitions des fumeurs, première tentative
ggplot(data_original, aes(x = factor(probleme_pulmonaire), fill = tabac)) +
  geom_bar(position = "fill") +
  labs(title = "Problèmes pulmonaires selon le score de tabac brut",
       x = "Problème pulmonaire", y = "Proportion", fill = "Score tabac") +
  theme_minimal()
|
```

Enfin, un autre graphique en barres empilées a été produit pour représenter l'exposition au tabagisme passif en fonction de la présence de problèmes pulmonaires. Ce graphique utilise la variable **problème_pulmonaire** sur l'axe des x, qui représente deux catégories : les individus sans problème pulmonaire (valeur 0) et ceux qui en présentent (valeur 1). Sur l'axe des y, il affiche la proportion d'individus exposés ou non au tabagisme passif dans chaque catégorie de santé pulmonaire. Les barres sont remplies selon les modalités de la variable **tabagisme_passif**, ce qui permet de comparer visuellement les groupes exposés à ceux non exposés. Le but de ce graphique est d'évaluer si l'exposition au tabagisme passif est associée

à une plus grande fréquence de problèmes respiratoires. En observant les barres, on constate que la proportion d'individus exposés est nettement plus élevée parmi ceux ayant des problèmes pulmonaires. Cela suggère un lien important entre cette exposition et l'état de santé pulmonaire, ce qui justifie l'intégration de cette variable comme facteur explicatif dans notre modélisation. (Voir annexe page n°1 graphique n°3)

3. Modélisation (2 Modèle différent)

La modélisation est réalisée à l'aide d'une régression logistique (fonction **glm()**), dont l'objectif est d'estimer la probabilité d'apparition d'un problème pulmonaire à partir d'un ensemble de variables explicatives. La variable cible **probleme_pulmonaire** est de nature binaire (0 = absence, 1 = présence), ce qui justifie naturellement l'emploi du modèle logit. Celui-ci est particulièrement adapté à la modélisation de variables qualitatives binaires, car il permet d'estimer directement l'effet marginal de chaque variable explicative en termes d'odds ratio

Dans un premier temps, un modèle complet a été estimé avec la commande suivante :

```
modele_logit <- glm(probleme_pulmonaire ~ age + sexe + tabac + tabagisme_passif + situation,
                      data = data_original, family = binomial)
```

Dans ce modèle, la variable **tabac** est incluse t elle quelle, avec toutes ses modalités numériques transformées en facteurs. Cela génère un très grand nombre de coefficients pour chaque valeur distincte du niveau de consommation. Cette approche, bien que complète, a conduit à des résultats peu lisibles, notamment dans les graphiques de visualisation des effets, car les intervalles de confiance associés à certaines modalités étaient très larges et les ordres de grandeur extrêmes

Afin de rendre l'interprétation plus accessible, un modèle simplifié a été mis en place. Dans ce second modèle, la variable **tabac** est remplacée par **classe_tabac**, une variable catégorielle construite à l'aide de la fonction **cut()** qui regroupe les niveaux de tabagisme en quatre classes : « Non-fumeur », « Faible », « Moyen » et « Fort »

```
modele_simplifie <- glm(probleme_pulmonaire ~ age + sexe + classe_tabac + tabagisme_passif + situation,
                          data = data_original, family = binomial)
```

Ce choix de regroupement améliore considérablement la stabilité du modèle et surtout la lisibilité graphique. L'effet de chaque variable a ensuite été analysé en termes d'odds ratios, avec calcul des intervalles de confiance à 95 % grâce au code suivant :

```
odds_ratios_simpl <- exp(coef(modele_simplifie))
or_ci_simpl <- exp(confint.default(modele_simplifie))
```

Un graphique de type dotplot avec barres d'erreur a été produit à l'aide de **ggplot2**, ce qui permet de visualiser pour chaque variable explicative l'intensité de son effet sur la probabilité d'avoir un problème pulmonaire, ainsi que la précision de cette estimation. La ligne rouge

horizontale au niveau de 1 représente l'absence d'effet : si l'intervalle de confiance d'une variable traverse cette ligne, son effet n'est pas considéré comme significatif. (Voir annexe page n°1 graphique n°4 & n°5)

Ce graphique permet d'observer que certaines modalités, notamment les niveaux de tabac les plus élevés (comme **tabac10** à **tabac14** dans le modèle initial), ont un effet très prononcé. En revanche, plusieurs variables s'avèrent non significatives dans ce modèle, comme la modalité **sexe et situation**.

Ces résultats sont obtenus via les intervalles de confiance calculés avec la fonction **confint.default()** : si un intervalle inclut la valeur 1, cela signifie que l'effet de la variable peut être nul. Le code suivant résume l'ensemble des effets sous forme de tableau :

```
or_ci_simpl <- exp(confint.default(model_e_simplifie))
```

```
effets_simpl <- data.frame(Variable = rownames(or_ci_simpl), OR = odds_ratios_simpl, IC_inf = or_ci_simpl[, 1], IC_sup = or_ci_simpl[, 2])
```

Ce tableau et le graphique associé permettent de trier visuellement les effets les plus significatifs. Dans le contexte de cette étude, on retient que les variables ayant un impact clair et significatif sont le niveau de tabagisme, le tabagisme passif, ainsi que dans une moindre mesure l'âge. Le sexe et la situation conjugale, bien que inclus dans le modèle pour contrôle, ne montrent pas d'effet direct significatif une fois les autres variables prises en compte

Les résultats montrent que l'âge a un effet croissant, le sexe masculin est associé à un risque accru, et les niveaux élevés de tabac sont fortement corrélés à la présence de problèmes. Le tabagisme passif est également un facteur aggravant.

*(l'introduction du modèle simplifié a donc permis de conserver la richesse explicative du modèle complet tout en rendant les effets interprétables et exploitables pour une analyse en santé publique obtenus sont transformés en odds ratios avec **exp(coef())**, ce qui permet une interprétation intuitive. Nous nous sommes beaucoup aidé des TP sur le site suivants “https://rpubs.com/tchalioui/TP_REG_LOG”)*

4. Evaluation du modèle

L'évaluation du modèle logit repose sur plusieurs critères complémentaires. Dans un premier temps, nous avons calculé le pseudo R² de McFadden à l'aide du package **pscl**, qui fournit une mesure de la qualité de l'ajustement du modèle . Le pseudo R² obtenu est satisfaisant et suggère que les variables explicatives contribuent de manière significative à la prédiction de la variable cible

Nous avons également utilisé un test de la vraisemblance pour comparer notre modèle complet à un modèle restreint sans variables. Ce test s'apparente à un test de Fisher global pour les modèles linéaires. Il est effectué via la commande suivante :

```
anova(modele_logit, test = "Chisq")
anova(modele_simplifie, test = "Chisq")
```

Ce test fournit une statistique du chi² avec une p-valeur associée. Une p-valeur inférieure à 0.05 permet de rejeter l'hypothèse selon laquelle les variables explicatives n'ont aucun effet, et donc de conclure que le modèle est globalement significatif

Enfin, nous avons mesuré la capacité discriminante du modèle à l'aide de la courbe ROC, générée avec le package **pROC**. Cette courbe compare les vrais positifs aux faux positifs pour différents seuils de probabilité, et l'aire sous la courbe (AUC) fournit une mesure globale de performance. Nous avons obtenu un AUC de 0.9879, ce qui indique une excellente performance. Voici le code utilisé :

```
proba <- predict(modele_logit, type = "response")
roc_curve <- roc(data_original$probleme_pulmonaire, proba)
plot(roc_curve, col = "darkblue", lwd = 2, main = "Courbe ROC du modèle logistique")
auc(roc_curve)

pr2(modele_simplifie)
proba_2 <- predict(modele_simplifie, type = "response")
roc_curve2 <- roc(data_original$probleme_pulmonaire, proba_2)
plot(roc_curve2, col = "darkred", lwd = 2, main = "Courbe ROC du modèle simplifié")
auc(roc_curve2)
```

Ce résultat signifie que le modèle a près de 99 % de chances de bien distinguer un individu malade d'un individu sain

L'évaluation du modèle par ces différents outils (pseudo R², test du chi², courbe ROC) montre qu'il est à la fois significatif, bien ajusté et performant dans sa capacité à prédire la présence de problèmes pulmonaires.

Un autre élément visuel essentiel à l'évaluation du modèle logit est le graphique représentant la distribution des probabilités prédictes par classe. Ce graphique repose sur les scores de prédiction générés par le modèle, c'est-à-dire les probabilités que chaque individu développe un problème pulmonaire. Ces probabilités sont ensuite croisées avec la variable observée **probleme_pulmonaire** pour visualiser leur répartition selon les deux groupes d'individus : ceux sans problème et ceux présentant un trouble pulmonaire (classe 1). (Voir annexe page n°2 graphique n°6)

L'axe horizontal du graphique affiche les probabilités prédictes comprises entre 0 et 1. L'axe vertical représente la fréquence d'apparition de ces probabilités dans l'échantillon. Chaque barre colorée traduit combien d'individus ont reçu un certain niveau de probabilité, en fonction de leur classe réelle. Les individus sains apparaissent en rose, tandis que les individus malades sont représentés en bleu clair.

L'interprétation de ce graphique est très éclairante. On y observe une séparation nette entre les deux groupes : la grande majorité des individus sans problème sont prédis avec une

probabilité très faible, proche de 0, tandis que les individus atteints sont concentrés vers les valeurs proches de 1. Cette nette différenciation confirme la qualité prédictive du modèle, qui parvient à bien distinguer les deux catégories de la variable cible.

Ce résultat visuel appuie les statistiques obtenues par la courbe ROC et l'AUC. Le fait que les deux classes soient bien séparées dans la distribution des probabilités témoigne de la puissance discriminante du modèle. Ainsi, ce graphique constitue une preuve complémentaire que le modèle logistique fonctionne de manière efficace et qu'il peut être considéré comme fiable dans le contexte de cette étude.

5. Effets variables

Afin d'interpréter correctement l'impact des variables du modèle logit simplifié, nous avons analysé les intervalles de confiance à 95 % des coefficients estimés à l'aide de la commande `confint.default(modele_simplifie)`

Une variable est considérée comme **statistiquement significative** si son **intervalle de confiance ne contient pas la valeur 0**, car cela signifie que son effet est suffisamment éloigné de l'hypothèse nulle (pas d'effet). À l'inverse, un intervalle incluant 0 indique que l'effet estimé pourrait ne pas être différent de zéro dans la population, et donc être attribuable au hasard.

Voici la synthèse corrigée des effets observés :

Variable	Significative ?	Intervalle IC	Interprétation
âge	Oui	[0.0586;02368]	Une augmentation de l'âge accroît la probabilité d'avoir un problème pulmonaire
sexé	Non	[1.0359;2.7468]	Le sexe ne semble pas influencer directement la santé pulmonaire
classe_TabacFaible	Non		Être fumeur faible n'a pas d'effet détectable statistiquement
classe_TabacMoyen	Oui	[1.3931;6.0433]	Être fumeur moyen augmente fortement le risque pulmonaire

classe_TabacFort	Non	[-7183;7233]	Fortement associée aux problèmes pulmonaires
tabagisme_passif	Non	[-7766;7729]	Estimation trop instablee
situation_encouple	Non	[-2.5774;1.5119]	aucun effet
situationmarie	Non	[-2.9323;1.6792]	aucun effet
situationveuf	Non	[-12751;12711]	aucun effet

Ce tableau met en évidence que seules deux variables, **âge** et **classe_TabacMoyen**, présentent un effet significatif avec des intervalles de confiance qui ne contiennent pas la valeur 0. Cela signifie que ces deux facteurs ont une influence statistiquement établie sur la probabilité de développer un problème pulmonaire dans notre échantillon. Le fait que l'âge soit significatif corrobore les connaissances médicales selon lesquelles le risque de pathologie respiratoire augmente avec l'âge. De même, le fait d'être un fumeur moyen (consommation modérée de tabac) augmente nettement le risque pulmonaire, ce qui est cohérent avec la littérature sur les effets nocifs du tabagisme.

En parallèle, les résultats de la commande summary(modele_simplifie) viennent confirmer cette lecture :

> `summary(modele_simplifie)`

Call:

```
glm(formula = probleme_pulmonaire ~ age + sexe + classe_tabac +
    tabagisme_passif + situation, family = binomial, data = data_original)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-9.26243	2.66398	-3.477	0.000507 ***
age	0.14776	0.04547	3.249	0.001157 **
sexehomme	0.85546	0.96503	0.886	0.375367
classe_tabacMoyen	3.71824	1.18628	3.134	0.001722 **
classe_tabacFort	24.95539	3677.72608	0.007	0.994586
tabagisme_passifTRUE	-18.64982	3953.16732	-0.005	0.996236
situationen_couple	-0.53277	1.04323	-0.511	0.609568
situationmarie	-0.62653	1.17644	-0.533	0.594336
situationveuf	-20.07124	6495.83578	-0.003	0.997535

les variables âge ($p = 0.001157$) et classe_tabacMoyen ($p = 0.001722$) présentent des valeurs p très faibles, ce qui indique une forte significativité statistique. À l'inverse, toutes les autres variables ont des valeurs p supérieures à 0.05, ce qui renforce notre décision de les considérer comme non significatives dans le cadre de cette analyse.

(*Pour approfondir la compréhension de la construction et de l'interprétation des intervalles de confiance sur R, nous précisons que nous avons utilisé les ressources suivantes pour comprendre et appliquer les commandes vue ci-dessus, nous nous sommes beaucoup aidées des vidéos suivantes "https://www.youtube.com/watch?v=2_z8BzmJRNC" et du lien suivants : "<https://datatab.fr/tutorial/confidence-interval>")

6. Conclusion

Cette étude statistique a permis de mettre en évidence les principaux facteurs de risque associés aux problèmes pulmonaires. Les résultats montrent clairement que l'âge et le tabagisme modéré (**classe_TabacMoyen**) sont les seuls déterminants statistiquement significatifs de la survenue de troubles respiratoires dans notre échantillon. En revanche, des variables souvent considérées comme importantes, telles que le tabagisme passif, le tabagisme fort ou encore le sexe, ne présentent pas d'effets significatifs. La régression logistique s'est avérée un outil efficace pour modéliser cette relation, avec un modèle performant validé par l'AUC très élevée. L'utilisation combinée de statistiques descriptives, d'analyses graphiques et d'interprétations des coefficients permet de fournir une lecture complète des comportements à risque.

Ces résultats peuvent alimenter des campagnes de prévention ciblant notamment les adultes exposés à une consommation régulière de tabac.

(Dans les différents scripts de brouillon produits au cours de cette étude, nous avons également exploré d'autres outils statistiques tels que l'Analyse en Composantes Principales (ACP), une matrice de corrélation, ainsi que plusieurs visualisations supplémentaires. Toutefois, ces éléments n'ont pas été intégrés dans le code final ni dans ce rapport, afin d'éviter de l'alourdir inutilement et de préserver la lisibilité et la cohérence de l'analyse principale.)

ANNEXE :



