Annales

Table des matières

1 2024

1.1 Statistiques élémentaires

1.1.1 Comment représenter graphiquement la distribution de la variable « niveau socio-économique » codée de la façon suivante : agriculteur, cadre, profession intermédiaire, commerçant, employé, ouvrier, sans emploi, autre.

- Le plus efficace = diagramme en bâtons, en prenant bien soin de présenter au mieux les différentes modalité pour que la lisibilité soit optimale et si besoin d'inclure une modalité « données manquantes »
- Camembert : possible mais peu recommandé, car moins lisible et moins précis qu'un diagramme en bâtons

1.1.2 Dans une étude épidémiologique vous mettez en évidence une corrélation entre l'IMC (indice de masse corporelle) et la tension artérielle moyenne égale à $r = 0{,}22$ avec un « p » égal à 0,0012. Un reviewer vous fait remarquer que le pourcentage de variance partagé par ces deux variables, égal à $0{,}222 = 0{,}0484$, est très faible et donc que cette relation est négligeable. Que lui répondez-vous ?

- La corrélation de Pearson (r) mesure la force et la direction d'une relation linéaire entre deux variables continues. Un r de 0,22 indique une corrélation positive faible entre l'IMC et la tension artérielle moyenne.
- Dans le cas d'une régression linéaire simple (donc à une seule variable explicative), le carré du coefficient de corrélation (r^2) représente le pourcentage de variance dans la variable dépendante (tension artérielle moyenne) qui peut être expliqué par la variable indépendante (IMC). Dans ce cas, un r^2 de 0,0484 signifie que seulement 4,84 % de la variance dans la tension artérielle moyenne peut être expliquée par l'IMC.
- Le p-value de 0,0012 indique que cette corrélation est statistiquement significative, ce qui signifie qu'il y a une faible probabilité que cette relation soit due au hasard.
- Le pourcentage de variance partagé (r^2) de 4,84 % indique que seulement une petite partie de la variance dans la tension artérielle moyenne peut être expliquée par l'IMC. Cependant, cela ne signifie pas nécessairement que la relation est négligeable. Même une faible corrélation peut être cliniquement significative, surtout si elle est cohérente avec d'autres recherches ou si elle a des implications pratiques importantes.
- Il est important de considérer le contexte clinique et les implications pratiques de cette relation, plutôt que de se concentrer uniquement sur la force de la corrélation ou le pourcentage de variance partagé.

Vraie correction :

L'interprétation de la force d'une association quand celle-ci est représentée par un coefficient de corrélation (de Pearson).

Il n'y a pas de consensus sur ce point dans la littérature. Dans certaines disciplines, comme en économétrie, voire en génétique dans le domaine biomédical, il est effectivement habituel de discuter en termes de pourcentage de variance expliquée ou partagée.

C'est cependant critiqué, notamment du fait qu'un pourcentage de variance expliqué dépend fortement :

- 1/ de l'échantillonnage de l'étude (un échantillon homogène conduira à une faible variance phénotypique et donc à des pourcentages de variance expliquée qui seront également faibles),
- 2/ de l'importance des erreurs de mesure et du bruit (quand ce dernier est important, puisque par définition il ne peut pas être expliqué les pourcentages de variance partagées seront faibles).

Le coefficient r lui-même n'est pas simple à interpréter, dans la littérature il est souvent considéré qu'un $r < 0,2$ est plutôt faible, mais il s'agit d'un point de vue purement subjectif.

Notamment parce que l'importance clinique et de santé publique de la relation va influer sur le caractère négligeable ou pas de cette dernière.

1.1.3 A quoi servent les tests statistiques ?

- En médecine (et en recherche biomédicale en général), les tests statistiques sont utilisés pour analyser des données et tirer des conclusions sur des populations à partir d'échantillons. Ils permettent de déterminer si les observations faites dans un échantillon sont suffisamment fortes pour être généralisées à une population plus large.
- Il s'agit d'un processus inférentiel, c'est à dire qu'on utilise les données d'un échantillon pour faire des inférences sur une population.
- Tout résultat tiré d'un échantillon présente une incertitude quant à sa généralisation à la population. Les tests statistiques aident à quantifier cette incertitude en fournissant des mesures telles que les p-values et les intervalles de confiance.

1.1.4 Comment vérifier les conditions de validité d'une régression linéaire ?

- Linéarité : La relation entre les variables indépendantes et dépendantes doit être linéaire. Cela peut être vérifié en examinant les graphiques de dispersion des résidus.
- Homoscédasticité : La variance des résidus doit être constante à travers toutes les valeurs des variables indépendantes. Cela peut être vérifié en traçant les résidus.
- Normalité des résidus : Les résidus doivent suivre une distribution normale. Cela peut être vérifié en utilisant des tests de normalité (comme le test de Shapiro-Wilk) ou en examinant les graphiques Q-Q des résidus.

- Indépendance des résidus : Les résidus doivent être indépendants les uns des autres, mais c'est beaucoup plus difficile à vérifier.

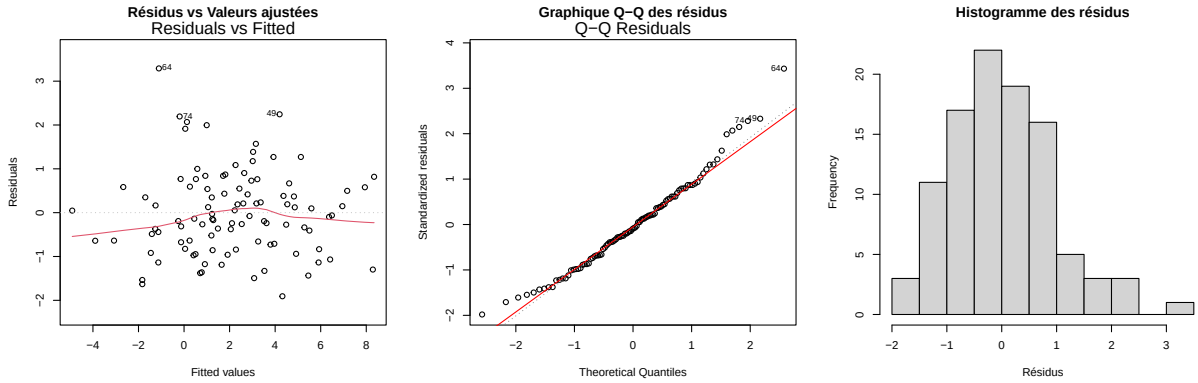


Figure 1: Vérification des conditions de validité d'une régression linéaire

1.2 Statistiques avancées

1.2.1 Vous souhaitez modéliser le nombre de passages aux urgences dans l'année pour des sujets issus d'une population gériatrique. Lors de la phase initiale de l'analyse vous constatez que cette variable présente une sur-dispersion. A quoi cela correspond-il et quelles sont les solutions que vous avez pour réaliser une telle modélisation ?

- Le nombre de passage aux urgences est une variable de comptage, souvent modélisée par une régression de Poisson.
- La distribution de Poisson est caractérisée par une moyenne égale à la variance (paramètre λ)
- La surdispersion survient lorsque la variance est supérieure à la moyenne.
- Prévisible ici, car pour un sujet donné, la survenue d'un évènement n'est pas indépendante de la survenue d'un autre évènement (ex : pathologie sous-jacente expliquant plusieurs passages aux urgences).
- 2 types de solutions :
  - Changement de modèle :
    - Quasi-Poisson : permet de modéliser la surdispersion en introduisant un paramètre de dispersion.
    - Modèle binomial négatif : une alternative à la régression de Poisson qui gère naturellement la surdispersion.
  - Approches robustes :

* Estimateur sandwich : ajuste les erreurs standards pour tenir compte de la surdispersion.
* Bootstrap : méthode de rééchantillonnage pour estimer les erreurs standards robustes.

Vraie correction : Il s'agit d'une variable qui peut être assimilée à un processus de comptage. Il y a surdispersion quand la variance est supérieure à la moyenne, une loi de Poisson ne peut donc pas être utilisée. Cette surdispersion était prévisible ici, elle survient quand, pour un sujet donné, la survenue d'un évènement n'est pas indépendante de la survenue d'un autre évènement. Si un sujet est atteint d'une pathologie sous-jacente (dénutrition, cancer, BPCO, etc.) ce sera le cas. Solution : classiquement on utilise un modèle de quasipoisson ou un modèle binomial négatif. De façon plus « technique » on peut utiliser un estimateur sandwich ou un bootstrap. De façon transgressive on pourrait utiliser un modèle linéaire avec un estimateur robuste devariance.

1.2.2 Dans une étude épidémiologique vous avez utilisé un questionnaire de qualité de vie de 18 items conduisant à un score global. Le questionnaire que vous avez envoyé aux sujets répondants a été mal imprimé et l'item n°13 est parfois difficile à lire, de ce fait la réponse est souvent manquante. Comment faites-vous pour analyser la variable « score de qualité de vie » ?

- Il s'agit d'une donnée manquante.
- Identifier le mécanisme de données manquantes (MCAR, MAR, MNAR).
- Ici, probablement MAR (Missing At Random) car la difficulté de lecture peut être liée à des caractéristiques observables (ex : âge, niveau d'éducation).
- Rappel :
  - MCAR (Missing Completely At Random) : les données manquantes sont indépendantes des valeurs observées et non observées (serait le cas si il manquait des données au hasard, par exemple si certaines pages du questionnaire étaient manquantes).
  - MAR (Missing At Random) : les données manquantes dépendent uniquement des valeurs observées.
  - MNAR (Missing Not At Random) : les données manquantes dépendent des valeurs non observées (on ne peut pas vérifier ce mécanisme avec les données disponibles).
- Dans la correction : c'est une MCAR !!! donc bizarre car c'est cohérent que la difficulté de lecture soit liée à des caractéristiques observables (ex : âge, niveau d'éducation).
- Imputation : globalement éviter l'imputation multiple qui va complexifier l'analyse du score global.
- On peut utiliser un échantillonneur de Gibbs dans le cas d'un questionnaire de qualité de vie car il permet d'imputer les valeurs manquantes en tenant compte de la structure des données et des relations entre les items du questionnaire.

1.3 Pourquoi le modèle d'analyse factorielle à 1 facteur est-il fondateur de la psychométrie (discipline qui a fondé les méthodes statistiques utilisées dans le domaine des mesures subjectives) ?

- Le caractère fondateur du modèle d'analyse factorielle à un facteur réside dans sa capacité à modéliser la relation entre des variables observées (items du questionnaire) et une variable latente (le facteur unique).
- Il s'agit d'un questionnaire **unidimensionnel**, c'est à dire que tous les items mesurent une seule et même dimension sous-jacente (le facteur unique).
- Mathématiquement, le modèle d'analyse factorielle à un facteur postule que chaque variable observée peut être exprimée comme une combinaison linéaire du facteur unique plus une erreur spécifique à chaque variable.
- Ce modèle permet d'estimer la charge factorielle de chaque item, qui indique dans quelle mesure chaque item est lié au facteur unique.

Vraie correction :

- Il y a un intérêt mathématique.
- Modèle AF1 : $X_i = a_i + b_i\Theta + \varepsilon_i$ justifie le fait que la somme des X_i est une estimation asymptotique (en fonction du nombre d'items n) de Θ, la variable d'intérêt sous-jacente.
- En effet, la somme des ε_i croit comme racine de n alors que la somme des $b_i\Theta$ croit comme n.
- Il y a une intérêt conceptuel. Le modèle AF1 suggère explicitement que chaque item estime le même concept Θ (en plus d'un bruit ε_i). Le score globale doit donc vraisemblablement bel et bien mesurer ce que nous espérons qu'il mesure.

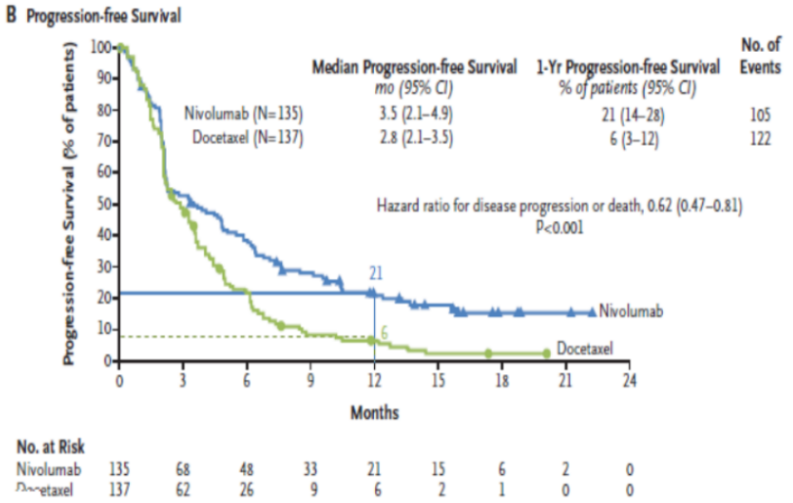
1.4 En quoi consiste l'analyse des données dans une étude qualitative ?

- Dans une étude qualitative (analyse d'un discours par exemple), l'analyse est faite par des sujets humains.
- Il s'agit d'une série successive de codage du texte à analyser par le codeur : du plus près du texte (micro-codage), au plus synthétique (codage axial ou codage macroscopique consistant en un regroupement des micro-codes).
- Pour améliorer la qualité du codage, plusieurs codeurs travaillent en parallèle (triangulation).
- Les dernières inclusions sont stoppées lorsqu'une saturation du codage est observée (donc le codage peut être fait au fur et à mesure des inclusions).

2 2023

2.1 Outils de la méthodologie

2.1.1 La comparaison de 2 anti-cancéreux conduit aux courbes de Kaplan- Meyer suivantes, avec un HR égal à 0,62. Un collègue vous fait remarquer que l'hypothèse des risques proportionnels n'est pas vérifiée. À quoi cela correspond-il et quelles en seraient les conséquences potentielles ?



- Sur la courbe, les deux courbes de survie se croisent, ce qui indique que le rapport des risques n'est pas constant dans le temps.
- L'hypothèse des risques proportionnels est une condition clé pour l'utilisation du modèle de Cox (pas pour la simple comparaison de Kaplan-Meier et le test du log-rank).
- Elle correspond à l'idée que le rapport des risques entre les deux groupes de traitement reste constant au fil du temps, c'est à dire que le rapport des probabilité d'évènement est constant au cours du temps (ce n'est pas le risque qui est constant, mais le rapport des risques).
- Ici, le rapport des risques n'est visiblement pas constant : il est proche de 1 les trois premiers mois, puis il diminue fortement par la suite.
- Les conséquences potentielles de la violation de cette hypothèse sont discutées dans la littérature. Certains considèrent que le HR (ici 0,62) n'est pas interprétable et qu'il faut utiliser un modèle avec une variable dépendante du temps. D'autres considèrent qu'il est interprétable mais comme un HR moyen, favorisant probablement les temps précoces par rapport aux temps tardifs.

2.1.2 Dans la partie « méthodologie » d'un article, les auteurs expliquent que comme il ne peuvent pas accepter l'hypothèse de mécanisme « MAR » de leurs données manquantes, ils ont été conduit à négliger ces dernières. Que pensez-vous de cette stratégie d'analyse ?

- Le mécanisme MAR (Missing At Random) suppose que la probabilité qu'une donnée soit manquante dépend uniquement des valeurs observées.
- Donc qu'on peut ignorer les données manquantes sans introduire de biais dans l'analyse car les données manquantes sont aléatoires par rapport aux valeurs non observées.
- Si les auteurs disent qu'ils ne peuvent pas dire que c'est MAR, cela suggère que les données manquantes pourraient être MNAR (Missing Not At Random), où la probabilité qu'une donnée soit manquante dépend des valeurs non observées donc ignorer les données manquantes pourrait introduire un biais significatif dans les résultats.

2.1.3 Quelles sont les circonstances où il est licite d'utiliser une méthode « pas à pas » (stepwise) de sélection de variables dans une régression logistique ?

- Les méthodes pas à pas (stepwise) de sélection de variables sont controversées en raison de leur tendance à surajuster les données et à produire des modèles instables.
- Dans l'idéal, il vaut mieux sélectionner les variables en se basant sur des connaissances préalables, la littérature existante, et des critères cliniques.
- A la limite, on peut réaliser une analyse univariée initiale pour identifier les variables à inclure selon les résultats univariés.

Correction : En pratique cette utilisation fait sens pour calculer des scores de prédiction. Il faut alors estimer le modèle sur un échantillon test puis apprécier ses performances prédictives sur un échantillon de validation. En effet, si l'on souhaite non pas prédire, mais plutôt rechercher des facteurs de risque susceptibles d'expliquer la survenue d'une maladie, alors le modèle doit être construit avant tout sur des bases cliniques ou physiopathologiques. Si ces dernières sont inexistantes, alors une régression pas à pas peut être envisagée, mais ici aussi avec un échantillon test et un échantillon de validation.

2.1.4 Dans la représentation graphique d'une matrice de corrélation par analyse en composantes principales, pourquoi les points variables sont-ils à l'intérieur d'un cercle ?

- Avant de réaliser l'ACP sur une matrice de corrélation, les variables sont centrées et réduites.
  - Après centration-réduction, chaque variable a une moyenne de 0.
  - Après centration-réduction, chaque variable a un écart-type de 1.
- Comme la variance = écart type^2, chaque variable a une variance de 1.

- L'ACP décompose la variance totale des variables en variance expliquée par chaque composante principale.
  - Les coordonnées d'une variable dans le plan des deux premières composantes principales (PC1, PC2) sont les corrélations de cette variable avec PC1 et PC2.
    * L'abscisse = corrélation avec PC1.
    * L'ordonnée = corrélation avec PC2.
    * Par définition, une corrélation de Pearson est comprise entre −1 et 1.
  - La distance d'une variable à l'origine dans ce plan dépend de la somme des carrés de ses corrélations avec PC1 et PC2.
  - La somme des carrés des corrélations d'une variable avec toutes les composantes principales est égale à la variance totale de cette variable.
- Ici, la variance totale vaut 1 (car les variables sont standardisées).
- Donc la somme des carrés des corrélations avec toutes les composantes principales vaut 1.
- Par Pythagore, la distance au carré à l'origine dans le plan (PC1, PC2) est égale à la somme des carrés des corrélations avec PC1 et PC2.
- Cette somme est toujours inférieure ou égale à la somme des carrés des corrélations avec toutes les composantes principales, qui vaut 1.
- Donc la distance à l'origine dans le plan (PC1, PC2) ne peut pas dépasser 1.

2.2 Statistiques élémentaires

2.2.1 Qu'est-ce qu'un beeswarm plot (diagramme en essaim d'abeilles) ? Quels en sont les intérêts et les limites ?

- Un beeswarm plot est un diagramme en essaim d'abeilles qui affiche chaque point de données individuel dans une distribution.
- La distribution est affichée car les points sont ajustés horizontalement pour éviter le chevauchement.
- C'est bien quand y a pas trop de données, sinon ça fait des trucs super larges.
- Visualisation des sous-groupes en faisant des couleurs différentes.

2.2.2 Pour le simple calcul d'un coefficient de corrélation entre 2 variables quantitatives dont les distributions sont toutes deux fortement asymétriques, faut-il privilégier le coefficient de corrélation de Spearman ou celui de Pearson ?

- La distribution normale n'est pas une condition nécessaire pour utiliser le coefficient de corrélation de Pearson.
- Les conditions nécessaires au calcul du coefficient de Pearson sont :
  - Relation linéaire entre les deux variables.
  - Homoscédasticité (variance constante des résidus).
- Si ces conditions sont remplies, Pearson est plus puissant que Spearman.
- Spearman : comme il est basé sur les rangs, il est moins sensible aux valeurs extrêmes et aux distributions asymétriques.

2.2.3 Un interne compare l'âge moyen des patientes hospitalisées en 2022 dans 2 services spécialisés dans la prise en charge de l'anorexie mentale. Il trouve 12 ans dans un cas et 15 ans dans l'autre et conclut qu'il y a une différence. Le chef de service l'interpelle en lui faisant remarquer que l'on ne peut affirmer une telle conclusion qu'après avoir réalisé un test statistique. Qu'en pensez-vous ?

- Effectivement, c'est différent !
- Il faut réaliser un test statistique pour savoir si c'est généralisable à une population infinie sous jacente.
- Il est licite de faire un test si la question posée est "est-ce que l'écart observé entre les deux moyennes est compatibe avec l'hypothèse que ces deux populations de patientes sont tirées au sort d'une population infinie sous-jacente ?"

2.2.4 Un modèle linéaire explique la variable Y « VEMS » (volume expiratoire maximum par seconde, en litres par seconde) en fonction de la variable X1 « âge » (en années) et de la variable X2 « statut tabagique » ('non-fumeur', 'ancien fumeur', 'fumeur actuel'). Comment le logiciel procède-t-il pour tester l'effet de X2 sur Y à X1 constant ?

- Il faut d'abord se demander si X2 est catégorielle pure (non ordonnée) ou ordonnée.
- Le modèle détermine une valeur de référence pour la variable et va tester l'effet des autres valeurs de la variable sur la moyenne du VEMS par rapport à la variable de référence.

- **Si X2 est catégorielle pure** : on crée deux variables binaires (dummy coding) avec “non-fumeur” comme référence, puis on teste conjointement que les deux coefficients sont nuls ($a'_1 = a'_2 = 0$). Donc le logiciel va calculer l’effet sur la moyenne du VEMS des fumeurs actuels et des anciens fumeurs par rapport aux non-fumeurs, à âge constant.
- **Si X2 est ordonnée** : on recodage simplement en (0, 1, 2) et on teste un seul coefficient ($a_2 = 0$).

3 2022

3.1 Outils de la méthodologie

3.1.1 A quoi est égal l'aire sous une courbe de Kaplan Meyer ?

- L'aire sous la courbe de Kaplan-Meier (AUC) n'est pas une mesure standard en analyse de survie.
- L'aire sous la courbe de Kaplan-Meier n'a pas d'interprétation directe en termes de probabilité ou de risque.
- S'il n'y a pas de censure, l'aire sous la courbe de Kaplan-Meier est égale à la durée moyenne de survie.

3.1.2 Vous publiez un article étudiant les facteurs de risque de refus de vaccination contre le COVID. Partant de 30 facteurs potentiel votre modèle final comprend 8 variables explicatives. Un reviewer vous demande d'appliquer une correction de Bonferroni sur vos résultats pour limiter le risque de « comparaisons multiples ». Que répondez-vous ?

- On se place ici dans une logique d'analyse statistique inférentielle selon Fisher.
- Il existe un article qui dit que dans une étude épidémiologique, la correction pour comparaisons multiples n'a pas de sens (To adjust, or not to adjust, for multiple comparisons, Hooper, J Clin Epidemiol. 2025).
- La correction de Bonferroni est très conservatrice et peut augmenter le risque d'erreurs de type II (faux négatifs).
- Il est préférable de se concentrer sur la cohérence des résultats avec la littérature existante et les connaissances cliniques.

3.1.3 Intérêts et inconvénients du modèle log-binomial.

- Le modèle log-binomial est utilisé pour estimer les risques relatifs dans les études épidémiologiques avec une variable dépendante binaire.
- Les avantages du modèle log-binomial :
  - Il fournit une estimation directe du risque relatif, ce qui est souvent plus interprétable que les odds ratios.
  - Il est approprié lorsque la prévalence de l'événement d'intérêt est élevée.
- Les inconvénients du modèle log-binomial :

  - Il peut rencontrer des problèmes de convergence, surtout lorsque les probabilités prédites sont proches de 1.
  - Il nécessite des ajustements spécifiques pour gérer les données corrélées ou les effets aléatoires.

3.1.4 Un collègue disposes d'une très petit échantillon (n=10). Les statistiques paramétriques étant difficilement utilisables il pense recourir à un bootstrap. Il vous demande conseil, que lui dites-vous ?

- Le bootstrap est une méthode de rééchantillonnage qui peut être utile pour estimer la variabilité des statistiques dans de petits échantillons.
- Cependant, avec un échantillon aussi petit (n=10), le bootstrap peut ne pas fournir des estimations fiables.
- Il est important de s'assurer que les hypothèses sous-jacentes du bootstrap sont respectées.
  - Le bootstrap repose sur l'hypothèse que l'échantillon est représentatif de la population.
  - Avec un échantillon très petit, cette hypothèse peut être difficile à vérifier.
- Pour le bootstrap, il faut au moins n=30.

3.2 Statistiques élémentaires

3.2.1 Vous calculez une corrélation entre un score de qualité de vie et une durée de consultation et trouvez un résultat r=0.34. Vos interlocuteurs cliniciens vous demandent ce que signifie concrètement cette valeur. Que répondez-vous ?

- Le coefficient de corrélation de Pearson (r) mesure la force et la direction d'une relation linéaire entre deux variables continues.
- Globalement, si r=1, les deux variables sont parfaitement corrélées positivement, si r=-1, elles sont parfaitement corrélées négativement, et si r=0, il n'y a pas de corrélation linéaire entre les deux variables.
- Mais il n'y a pas de seuil entre tout ça pour dire si c'est faible, modéré ou fort.
- Dans le cas d'une régression linéaire simple, le carré du coefficient de corrélation (r^2) représente le pourcentage de variance dans la variable dépendante (durée de consultation) qui peut être expliqué par la variable indépendante (score de qualité de vie).
- Mais on sait pas vraiment comment interpréter le pourcentage de variance partagée.

3.2.2 Vous souhaitez comparer les moyennes de deux séries de 15 mesures. Les conditions de validité du test t de Student imposent alors de vérifier une normalité distributionnelle. Que pensez-vous du recours à un test de normalité dans une telle situation ?

- Les tests de normalité ne sont pas vraiment utiles car il repose sur des hypothèses qui sont difficiles à vérifier avec un petit échantillon.
- Les tests s'interprètent comme "non significatif" = "normalité vérifiée", mais avec un petit échantillon, la puissance du test est faible.
- Comme la puissance est faible et impossible à estimer, on ne peut pas vraiment se fier aux résultats du test.
- Il est préférable d'utiliser des méthodes graphiques (histogrammes, Q-Q plots) pour évaluer la normalité.

3.2.3 Il est parfois dit que l'on ne peut pas accepter une hypothèse nulle, on peut simplement ne pas la rejeter, qu'en pensez-vous ?

- En statistique de Fisher : non ! on ne peut pas accepter l'hypothèse nulle, on peut seulement ne pas la rejeter (car on ne conclut pas grand chose quand la *p-value* est élevée)
- En statistique de Neyman-Pearson : si ! on peut accepter l'hypothèse nulle si on a un test avec une puissance suffisante car on a défini des risques d'erreurs de type I et II a priori.

3.3 Une étude européenne sur la pratique du sport est réalisée dans 10 pays. Un modèle ajusté sur plusieurs covariables est réalisé et la variable « pays » est introduite comme effet (fixe). Si l'on est embarrassé par le choix d'un pays de référence (pouvant être perçu comme arbitraire et stigmatisant), comment procéder pour obtenir tout de même un effet propre à chaque pays ?

- Par défaut, les pays sont transformés en 9 variables binaires avec une variable de référence.
- Pour éviter de choisir un pays de référence, on peut utiliser le codage "contrastes de Helmert" ou "contrastes de somme".
- En gros, il faut que les pays soient codés en variables binaires mais sans variable de référence
- Pour ça : un pays doit être codé -1 pour toutes les variables binaires

En tableau :

Pays	Pays 1	Pays 2	Pays 3	Pays 4	Pays 5	Pays 6	Pays 7	Pays 8	Pays 9	Pays 10
X1	1	0	0	0	0	0	0	0	0	-1
X2	0	1	0	0	0	0	0	0	0	-1
X3	0	0	1	0	0	0	0	0	0	-1
X4	0	0	0	1	0	0	0	0	0	-1
X5	0	0	0	0	1	0	0	0	0	-1
X6	0	0	0	0	0	1	0	0	0	-1
X7	0	0	0	0	0	0	1	1	0	-1
X8	0	0	0	0	0	0	0	1	0	-1
X9	0	0	0	0	0	0	0	0	1	-1

- Chaque coefficient estimé dans le modèle représente la différence entre le pays correspondant et la moyenne des autres pays.
- Cela permet d'obtenir un effet propre à chaque pays sans avoir à choisir un pays de référence spécifique.

4 2021

4.1 Outils de la méthodologie

4.1.1 Lors de la validation d'un instrument de mesure subjective pourquoi doit-on déterminer son niveau d'unidimensionnalité ?

- La validité d'un instrument de mesure subjective passe par plusieurs étapes :
  1. Que mesure t-il ? (validité de contenu)
  2. Comment le mesure t-il ? (validité de construit)
  3. Est-ce qu'il le mesure bien ? (fidélité)
- L'unidimensionnalité est une condition nécessaire pour que l'instrument mesure un seul et même construit.
- C'est une condition nécessaire à la réalisation d'un alpha de Cronbach, qui validera la consistence des items entre eux, c'est à dire la fidélité de l'instrument et son absence de redondance.

4.1.2 Comment apprécier l'accord inter-juge de 2 cliniciens portant le diagnostic oui/non de rhumatisme psoriasique chez 80 patients vus en consultation spécialisée de rhumatologie ? Comment expliquez-vous les résultats aux rhumatologues ?

- Il s'agit d'évaluer l'accord inter-juge sur une variable catégorielle, ici binaire (oui/non)
- On calcule donc le κ de Cohen, égal à $\frac{textconcordanceobserve - \text{concordance due au hasard}}{1 - \text{concordance due au hasard}}$
- Le κ varie entre -1 et 1, où 1 indique un accord parfait, 0 indique un accord équivalent au hasard, et des valeurs négatives indiquent un désaccord systématique.

4.1.3 Trouvez un exemple où 2 courbes de survies ne vérifient pas l'hypothèse des risques proportionnels. Comment les comparer dans un tel cas ?

- Exemple : deux courbes de survie qui se croisent, indiquant que le risque relatif entre les deux groupes change au fil du temps.
- Pour comparer ces courbes, c'est difficile car on ne peut pas appliquer de modèle de Cox ici du fait de l'absence ce proportionnalité des ratio de risque.
- On peut tout de même utiliser le test du log-rank, qui compare les courbes de survie globalement, même si l'hypothèse des risques proportionnels n'est pas vérifiée.
- NB : l'hypothèse de proportionnalité des ratio de risque signifie qu'à tout moment, le rapport entre les risques de deux groupes reste constant dans le temps.

4.1.4 Une régression logistique « pas à pas » (stepwise) peut être utilisée par un chercheur travaillant dans le domaine des sciences des données (datasciences) mais ne devrait pas être utilisée par un chercheur en épidémiologie. Pourquoi ?

- Régression stepwise = sélection automatique des variables explicatives en ajoutant ou en retirant des variables selon des critères statistiques (ex : p-value).
- En datascience : l'objectif de la construction d'un modèle est de PRÉDIRE la survenue d'un évènement (ex : maladie), donc ça a du sens car on veut le meilleur modèle possible (encore qu'il y a un risque de surapprentissage)
- En épidémiologie : l'objectif de la construction d'un modèle est d'EXPLIQUER la survenue d'un évènement (ex : maladie), donc il faut sélectionner les variables explicatives en se basant sur des connaissances cliniques et physiopathologiques. Le risque de biais est élevé avec une régression stepwise.

4.2 Statistiques élémentaires

4.2.1 Quels sont les avantages et les inconvénients d'un histogramme pour représenter la distribution d'une variable quantitative ?

- Avantages :
  - Interprétation simple
  - Permet de voir la distribution
  - Permet de déterminer le ou les modes
- Inconvénients :
  - Choix arbitraire des 'breaks"
  - Effet de bord = c'est à dire que selon où on place les `breaks`, la forme de l'histogramme peut changer

4.2.2 Quand est-il utile de calculer un coefficient de corrélation de Spearman à la place du traditionnel coefficient de corrélation de Pearson ?

- Le coefficient de corrélation de Pearson mesure la force et la direction d'une relation linéaire entre deux variables continues.
- Le coefficient de corrélation de Spearman mesure la force et la direction d'une relation non linéaire entre deux variables continues ou ordinales, et est basé sur les rangs des données.
- Si les données ne suivent pas une distribution normale, un Spearman est préférable.
- Mais en vrai, on pourrait faire un bootstrap pour rendre les variables normales puis faire un coefficient de Pearson.

4.2.3 Un collègue vient vous voir et vous dit avec enthousiasme : « ça marche, j'ai un « p » à 0,012, je vais pouvoir publier mon papier, c'est super !». Que lui répondez-vous ?

Bravo ! Mais en fait un calcul de « p » n'a pas grand-chose à voir avec le fait que l'on publie ou pas un papier…

La publication du papier c'est avant tout une question scientifique sous-jacente, des données, des résultats, des interprétations. Le « p » n'est qu'une étape de ce processus qui permet d'y voir un peu clair. A ce sujet, a-t-il écrit au préalable un plan d'analyse statistique ?

Si, comme c'est souvent le cas, il n'y en avait pas, alors il faut faire preuve d'une très grande humilité dans l'interprétation de ce « p » (nous, statisticiens, diront que nous sommes vraisemblablement dans une inférence inductive comme le proposait Fisher).

Dans tous les cas, un « p » ne prouve rien à lui tout seul. Il permet juste de donner au lecteur une certaine idée du rôle possible des fluctuations d'échantillonnage dans le résultat trouvé.

4.2.4 Si l'on souhaite inclure la variable « niveau socio-économique » dans un modèle de régression linéaire qu'est-ce que cela implique numériquement pour l'estimation du modèle ?

- Le niveau socio-économique est une variable catégorielle.
- On a deux options :
  - soit on l'estime avec un niveau de référence, donc on crée k-1 variables binaires (dummy coding) si on a k niveaux.
  - soit on l'estime avec des contrastes (contrastes de Helmert ou de somme) pour éviter de choisir un niveau de référence, dans ce cas la somme des coefficients doit être nulle et la différence sera faite par rapport à la moyenne des autres niveaux.

5 2019

5.1 Outils de la méthodologie