Projet de Statistique (partie Tests d'hypothèses)

Contexte : la capacité pour un couple à avoir un enfant est une préoccupation majeure des individus et un élément important de la santé des populations. En France, un couple sur sept consultera au cours de sa vie reproductive pour des difficultés à concevoir. Certaines études mettent en évidence une relation entre surpoids/obésité et infertilité chez l'homme, ce qui dénoterait d'un impact de la qualité de vie sur la fertilité de l'homme.
Le fichier `fertiproj.txt` contient des données de couples venus consulter pour infécondité masculine entre 2000 et 2004 au Centre de Stérilité Masculine (CSM) à Toulouse. Les patients ont réalisé un spermogramme au cours de leur première consultation (cf. wikipedia pour plus de détails sur le spermogramme).
L'objectif de l'étude est de comprendre les liaisons qui peuvent exister entre les différentes variables et d'identifier en particulier l'impact des habitudes de vie (tabac, alcool, IMC) sur les éléments du spermogramme (volume, concentration, numération totale de spermatozoïdes, mobilité et vitalité), qui sont liés au problème d'infertilité.
L'Organisation Mondiale de la Santé (OMS) a défini un indice, appelé Indice de Masse Corporelle (IMC) comme le standard pour évaluer les risques liés au surpoids. Il se calcule en faisant le rapport du poids (en kg) sur la taille (en m) élevée au carré. L'OMS a également défini des intervalles standard en se basant sur la relation constatée statistiquement entre l'IMC et le taux de mortalité. Ces intervalles sont les suivants :

- IMC $<$ 18,5 : maigreur
- 18,5 $\leq$ IMC $<$ 25 : corpulence normale
- 25 $\leq$ IMC $<$ 30 : surpoids
- IMC $\geq$ 30 : obésité

Vous allez travailler sur les variables suivantes :

- $alcool$: 1 si le patient est alcoolique (boit plus d'un litre d'alcool par jour), 0 sinon,
- $tabagisme$: 1 si le patient fume, 0 sinon,
- vol : volume de l'éjaculat (en ml),
- $conc$: concentration en spermatozoïdes de l'éjaculat (en millions par ml),
- vit : vitalité, c'est-à-dire pourcentage de spermatozoïdes vivants dans l'éjaculat,
- $exam$: nombre d'examens anormaux à la consultation,
- age : âge du patient lors de la consultation (en années),
- $dureeinf$: durée d'infertilité (en mois), c'est-à-dire durée pendant laquelle le couple a essayé en vain d'avoir un enfant,
- imc : indice de masse corporelle (en kg/m^2),
- mob : mobilité, c'est-à-dire pourcentage de spermatozoïdes mobiles dans l'éjaculat,
- $fievre$: présence d'une forte fièvre moins de trois mois avant la consultation (la fièvre a un impact sur la spermatogénèse et peut la perturber).

1. Faire une introduction en décrivant la population concernée par l'étude et préciser les variables étudiées et le but de l'étude.
2. Effectuer les créations de variables suivantes avec `R`. Sauf mention contraire, ce sont ces nouvelles variables qui seront utilisées dans la suite du projet.
   (a) Recoder les variables $alcool$ et $tababisme$ en oui / non au lieu de 1 / 0.
   (b) Créer la variable num (pour numération), qui correspond au nombre de spermatozoïdes présents dans l'éjaculat (en millions). Elle s'obtient en faisant le produit de la concentration par le volume.
   (c) Créer la variable $imcc$, correspondant à l'IMC en 3 classes, en vous basant sur les classes définies par l'OMS, mais en regroupant les individus maigres, trop peu nombreux dans l'échantillon, avec les individus de corpulence normale.
   (d) Créer la variable $examc$ en 2 classes : "aucun examen anormal" et "au moins un examen anormal".
3. Etude descriptive

   Commencer l'étude statistique par une étude descriptive de toutes les variables du fichier initial (sauf imc et $exam$ qui ne serviront plus) ainsi que des nouvelles variables créées. Pour chaque variable, donner son type (quantitatif continu, quantitatif discret ou qualitatif) et décrire sa distribution à l'aide des outils appropriés (graphiques, résumés numériques et commentaires). Tester la normalité de la variable vit.
   Rappel : toute sortie doit être commentée.

4. Comparaison de la population d'où est extrait l'échantillon avec la population française en 2002.

   (a) Il est souvent conseillé d'attendre deux ans d'essais infructueux avant de consulter pour infertilité. Est-ce que les couples **de votre échantillon** ont suivi ce conseil ? **(Pour cette question uniquement, on ne demande pas de faire de test.)**

   (b) En France, en 2002, la répartition de l'IMC en classes était la suivante : 58,4 % d'individus maigres ou normaux, 29,2 % d'individus en surpoids, et 12,4 % d'individus obèses. Peut-on considérer que la population que vous étudiez présente la même répartition de l'IMC en classes ? Si non, préciser.

   (c) En 2002, la France comptait 36% d'hommes fumeurs. La population que vous étudiez fume-t-elle autant ?

   (d) Un éjaculat compte en moyenne 250 millions de spermatozoïdes. Les individus qui consultent pour infertilité au CSM ont-il une numération en-dessous de cette moyenne (ce qui pourrait expliquer leur problème de fécondité) ?

5. Influence de l'âge sur les autres variables

   (a) Le volume de l'éjaculat diminue-t-il avec l'âge ?

   (b) L'âge a-t-il une influence sur la vitalité des spermatozoïdes, sur leur mobilité ?

   (c) Le comportement tabagique dépend-il de l'âge ?

   (d) La corpulence augmente-t-elle avec l'âge (faire cette question avec l'IMC en classes) ? Faire si besoin des tests de comparaisons multiples.

6. Influence des addictions et des conditions de vie

   (a) Y-a-t-il un lien entre l'addiction au tabac et l'addiction à l'alcool ?

   (b) La corpulence (en classes) est-elle liée au statut tabagique ?

   (c) La vitalité des spermatozoïdes est-elle perturbée par l'abus d'alcool ? Par le fait de fumer ?

   (d) Un accès de fièvre récent a-t-il un effet sur la vitalité de l'éjaculat ?

   (e) Un accès de fièvre récent peut-il expliquer la présence d'au moins un examen anormal ?

   (f) Quel lien existe-t-il entre la numération de l'éjaculat et la corpulence (faire cette question avec l'IMC en classes) ? Faire si besoin des tests de comparaisons multiples.

7. **Question bonus.** Faire une régression linéaire multiple pour évaluer l'impact des variables qualitatives du fichier sur la vitalité, en ajustant sur l'âge, le volume et la concentration. Commenter.

8. Conclusion de l'étude

   Faire une courte synthèse sous la forme d'un tableau récapitulatif montrant ce que vous avez compris des relations entre les différentes variables et en particulier des variables qui peuvent influer sur la vitalité.