MATH0487-2 - Éléments de statistiques

Feedback et solution Q4 Projet

---

1 Feedback

Majoritairement, le projet est très bien fait. Voici quelques points qui reviennent dans plusieurs projets.

- Pour l'étude de la population, il faut utiliser l'écart-type classique, non corrigé. Ce n'est pas celui par défaut de Matlab.
- Le coefficient de corrélation utilisé ici évalue la corrélation linéaire. Le fait qu'il soit faible ne signifie pas forcément que les deux variables sont indépendantes, mais plutôt que leur relation ne peut pas être représentée par une droite.
- C'est mieux de justifier, quand cela a été vu au cours, des résultats expérimentaux avec la théorie. Par exemple, pour la Q2b, il fallait mentionner le théorème central limite ou le fait que l'espérance de l'écart-type est biaisée (par exemple via l'inégalité de Jensen si s_{n-1} a été utilisé).
- Pour la question Q3(d), les deux types d'intervalles, Gauss et Student, permettaient de confirmer que la variable parente suit une loi normale. Pour construire un intervalle de confiance avec la loi de Student, la variable parente $\mathcal{X}$ ne suit pas une loi de Student, c'est le fait qu'on utilise s_{n-1} qui fait que l'on utilise plutôt une loi de Student, car $\frac{m_x-\mu}{s_{n-1}/\sqrt{n}}$ suit une loi de Student à (n-1) degrés de liberté lorsque $\mathcal{X} \sim \mathcal{N}(\mu, \sigma^2)$.
- Pour le test d'hypothèse, il était important de préciser les hypothèses, justifier les raisonnements et décrire la méthodologie.
- Concernant le code, attention à bien respecter les spécifications. Par exemple, on demandait un vecteur en entrée pour la fonction sum_up.m et plusieurs groupes ont considéré que l'entrée était le fichier data complet. Comme les codes sont testés avec un vecteur, ces fonctions ont renvoyé une erreur.

2 Solution Q4

Le but de cette question est de vous familiariser avec les tests d'hypothèses et de bien comprendre les différents types d'erreurs qui peuvent en résulter.

Le test d'hypothèse demandé consiste à comparer la moyenne des frais hospitaliers de deux populations distinctes, les fumeurs et les non-fumeurs.

L'hypothèse testée ici est : la différence entre les frais hospitaliers moyens des fumeurs et les frais hospitaliers moyens des non-fumeurs est égale à x.

La valeur de x s'obtient en calculant la différence entre les frais moyens des fumeurs de votre population et les frais moyens des non-fumeurs de votre population. Vous savez

donc, puisque vous avez accès à la population entière (différente pour chaque groupe), que l'hypothèse nulle est vraie.

Lorsque vous faites un test d'hypothèse, il faut toujours définir H0 et H1. Ici, pour l'hypothèse alternative, vous aviez le choix entre un test unilatéral ou bilatéral puisque celle-ci n'était pas précisée. Dans cette solution, nous présentons le test bilatéral.

2.1 Solution pour un test bilatéral - Q4a

Nous introduisons les notations suivantes pour la suite. Soit $\mathcal{F}_f$ la VA représentant les frais hospitaliers d'un fumeur et $\mathcal{F}_{nf}$ la VA représentant les frais hospitaliers d'un non-fumeur.

H0 : $\mu_{\mathcal{F}_f}$-$\mu_{\mathcal{F}_{nf}} = x$
H1 : $\mu_{\mathcal{F}_f}$-$\mu_{\mathcal{F}_{nf}} \neq x$

On génèrera pour le test deux échantillons iid de taille 50, un tiré parmi les fumeurs et un tiré parmi les non-fumeurs.

On utilise comme statistique pour le test la différence entre les moyennes de nos deux échantillons, $\Delta m = m_{\mathcal{F}_f}$-$m_{\mathcal{F}_{nf}}$. Par le théorème central limite et vu que $n > 30$, on suppose que $m_{\mathcal{F}_f}$ et $m_{\mathcal{F}_{nf}}$ suivent une loi normale [1].

On a alors $m_{\mathcal{F}_f} \sim \mathcal{N}\left(\mu_{\mathcal{F}_f}, \frac{\sigma^2_{\mathcal{F}_f}}{n_f}\right)$ et $m_{\mathcal{F}_{nf}} \sim \mathcal{N}\left(\mu_{\mathcal{F}_{nf}}, \frac{\sigma^2_{\mathcal{F}_{nf}}}{n_{nf}}\right)$, où n_f et n_{nf} représentent respectivement la taille d'échantillon des fumeurs et des non-fumeurs et sont égaux à 50 et $\sigma^2_{\mathcal{F}_f}$ et $\sigma^2_{\mathcal{F}_{nf}}$ [2] représentent la variance des frais des fumeurs et des non-fumeurs.

Donc, par la propriété d'additivité des gaussiennes et comme $m_{\mathcal{F}_f}$ est indépendante de $m_{\mathcal{F}_{nf}}$, Δm suit une loi normale. On a aussi

$E\{\Delta m\} = E\{m_{\mathcal{F}_f} - m_{\mathcal{F}_{nf}}\} = \mu_{\mathcal{F}_f} - \mu_{\mathcal{F}_{nf}}$ (par linéarité de l'espérance),

$V\{\Delta m\} = V\{m_{\mathcal{F}_f} + (-1)m_{\mathcal{F}_{nf}}\} = V\{m_{\mathcal{F}_f}\} + (-1)^2 V\{m_{\mathcal{F}_{nf}}\} = \frac{\sigma^2_{\mathcal{F}_f}}{n_f} + \frac{\sigma^2_{\mathcal{F}_{nf}}}{n_{nf}}$
(car $m_{\mathcal{F}_f}$ et $m_{\mathcal{F}_{nf}}$ sont indépendants).

On a donc $\Delta m \sim \mathcal{N}\left(\mu_{\mathcal{F}_f} - \mu_{\mathcal{F}_{nf}}, \frac{\sigma^2_{\mathcal{F}_f}}{n_f} + \frac{\sigma^2_{\mathcal{F}_{nf}}}{n_{nf}}\right)$.

Sous H0, $\Delta m \sim \mathcal{N}\left(x, \frac{\sigma^2_{\mathcal{F}_f} + \sigma^2_{\mathcal{F}_{nf}}}{50}\right)$.

On calcule alors la région critique, telle que si Δm se trouve dans cette région, on rejette H0.

Pour un test bilatéral, on rejette H0 si

$$\Delta m \notin \left[x - u_{1-\alpha/2}\frac{\sigma^2_{\mathcal{F}_f} + \sigma^2_{\mathcal{F}_{nf}}}{50}, x + u_{1-\alpha/2}\frac{\sigma^2_{\mathcal{F}_f} + \sigma^2_{\mathcal{F}_{nf}}}{50}\right],$$

1. C'est bien la moyenne d'échantillon qui suit une loi normale et pas la variable parente, comme on l'a vu aux questions 1 et 2.

2. Vous pouviez également considérer cette donnée comme inconnue et calculer s_{n-1} pour chaque échantillon.

où $u_{1-\alpha/2} = 1.96$, vu que $\alpha = 0.05$.

On réalise ensuite ce test 100 fois. Vous deviez trouver en moyenne un nombre de rejets proche de 5. Ce nombre proche de α est celui attendu, puisque α représente la probabilité de rejeter H0 alors qu'elle est vraie (ce qui est le cas ici). On peut également conclure que l'hypothèse d'une loi normale est justifiée.

Remarque : plusieurs étudiants ont écrit que comme le nombre de rejets est proche de α, H0 est vérifiée. Ce ne sera pas toujours le cas, vous pourriez avoir un test où la puissance du test $1 - \beta$ est très faible et est proche de 5% également.

2.2 Solution pour un test bilatéral - Q4b

Le raisonnement est similaire au point précédent. Les seules différences sont les échantillons et la variance de Δm qui doit être recalculée avec les personnes de plus de 50 ans. La région critique est donc différente entre la Q4a et la Q4b.

Ici, dans le cas du test bilatéral, on se trouve dans la situation où H1 est vraie. Donc lorsque vous analysez la proportion de rejet de l'hypothèse nulle, ce n'est plus α que vous estimez mais la puissance du test $1 - \beta$, qui représente la probabilité de rejeter H0 alors qu'elle est effectivement fausse. On obtient la même conclusion pour un test unilatéral à droite.

Par contre, dans le cas d'un test unilatéral gauche, H0 reste vraie et donc vous devriez trouver un pourcentage de rejet proche de α.