**Instructions:**

- Répondez aux questions suivantes à l'aide de SAS ou de R et fournissez le code utilisé pour vos analyses dans un fichier (`.R` ou `.sas`).
- Votre rapport doit être remis en ligne en format PDF et en version papier en classe et ne devrait pas faire plus de 10 pages; toute page excédentaire sera ignorée lors de la correction. Soyez concis mais précis: n'incluez que les sorties pertinentes.
- Interprétez vos résultats dans le contexte du problème et justifiez votre choix de tests d'hypothèse.
- Les erreurs sont pénalisées même si elles ne sont pas en lien direct avec la question.

**Contexte:** Les données `renfe` contiennent des informations sur 10 000 billets de trains vendus par la compagnie Renfe, l'entreprise ferroviaire publique espagnole. Les données incluent les variables:

- `prix`: prix du billet (en euros);
- `dest`: indicateur binaire du trajet, soit de Barcelone vers Madrid (`0`) ou de Madrid vers Barcelone (`1`);
- `tarif`: variable catégorielle indiquant le tarif du billet, un parmi `AdultoIda`, `Promo` et `Flexible`;
- `classe`: classe du billet, soit `Preferente`, `Turista`, `TuristaPlus` ou `TuristaSolo`;
- `type`: variable catégorielle indiquant le type de train, soit Alta Velocidad Española (`AVE`), soit Alta Velocidad Española conjointement avec TGV (un partenariat entre la SNCF et Renfe pour les trains à destination ou en provenance de Toulouse) `AVE-TGV`, soit les trains régionaux `REXPRESS`; seuls les trains étiquetés `AVE` ou `AVE-TGV` sont des trains à grande vitesse.
- `duree`: longueur annoncée du trajet (en minutes);
- `jour` entier indiquant le jour de la semaine du départ allant de dimanche (`1`) à samedi (`7`).

1. Faites une analyse exploratoire des données `renfe` afin
   - d'évaluer graphiquement les facteurs déterminant le prix et le temps de parcours.
   - de déterminer les caractéristiques distinctives des types de train.
   - d'établir s'il y a des différence entre les différents tarifs.

   Votre résumé devrait brosser un portrait complet de la situation. De plus, votre analyse devrait inclure des justificatifs (statistiques descriptives, tableaux de fréquences, histogrammes et boîtes à moustaches, etc. en plus d'informations supplémentaires tirées du site internet de la compagnie) pour étayer vos arguments.
2. Supposez qu'un analyste décide de ne conserver que les 1000 premières observations. Est-ce que ce serait un bon échantillon? Justifiez votre réponse en vous basant sur les informations disponibles dans la base de données.
3. On considère le temps de parcous pour les trains à grande vitesse (`AVE` et `AVE-TGV`). Le temps médian entre les deux villes dans la « population » est de $\nu = 2.833$ heures, tandis que la moyenne de la « population » est de $\mu = 2.845$ heures; ces quantités ont été déterminées sur la base des données complètes contenant plus de 2.3 millions d'entrées et sont donc considérées comme connues, contrairement à la plupart des applications pratiques.

   Une étude de simulation a été conduite pour déterminer le comportement de tests pour un échantillon dans un contexte d'échantillonage répété. L'algorithme suivant a été répété 10 000 fois:
   (a) sélection d'un sous-échantillon de taile $n = 100$.
   (b) calcul de la statistique du test-$t$ pour un échantillon correspondant à $\mathscr{H}_0 : \mu = \mu_0$ (versus $\mathscr{H}_0 : \mu \neq \mu_0$) pour différentes valeurs de $\mu_0$.
   (c) calcul de la statistique du test des signes pour le test bilatéral $\mathscr{H}_0 : \nu = \nu_0$ pour différentes valeurs de $\nu_0$.
   (d) calcul de la statistique du test des rangs signés de Wilcoxon pour le test bilatéral $\mathscr{H}_0 : \nu = \nu_0$ pour différentes valeurs de $\nu_0$.
   (e) sauvegarde des valeurs-$p$ associées à chacun des trois tests.

   Notez que le test des signes et le test de Wilcoxon sont deux tests pour la **médiane**.

   La fig. 1 montre le pourcentage de valeur-$p$ parmi les 10 000 qui sont plus petites que 0.05, c'est-à-dire la proportion de rejet (à un niveau de 5%) de $\mathscr{H}_0 : \mu = \mu_0$ contre l'alternative bilatérale à $\mu_0 \in \{2.83, \nu, 2.835, 2.84, \ldots, 2.995, 3\}$ (pour le test des signes et de Wilcoxon, nous testons si la médiane est égale à ces même valeurs). Utilisez la courbe de



Compiled 30/01/2020 at 10:12

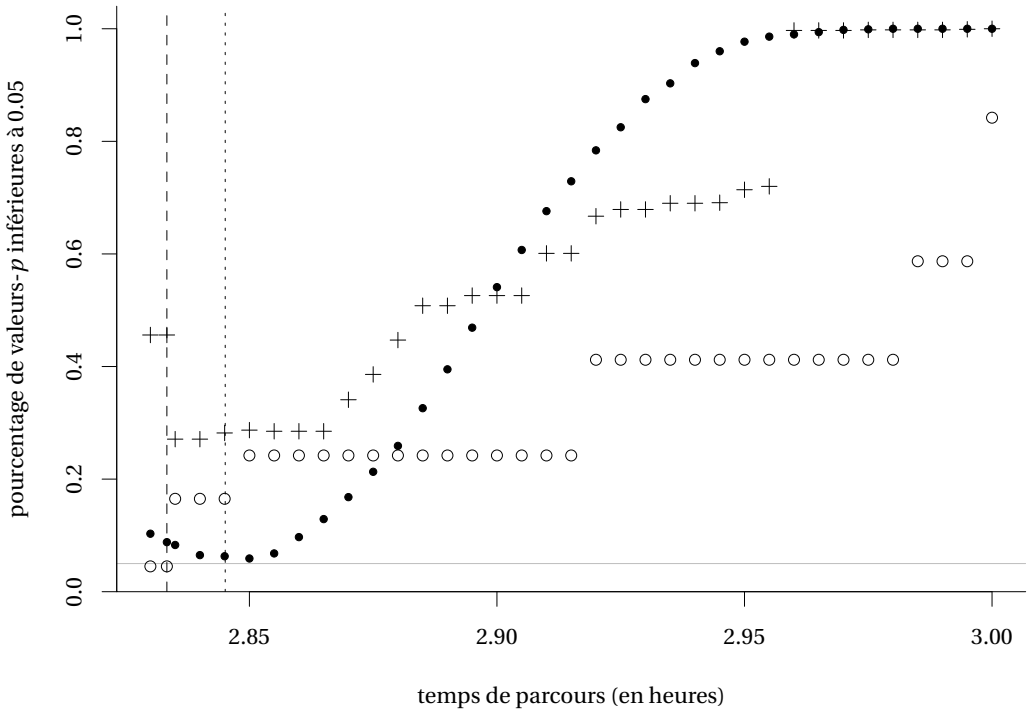


Figure 1: Courbe de puissance pour trois tests de localisation, soit le test-$t$ pour un échantillon (disque), le test des rangs signés de Wilcoxon (croix) et le test des signes (cercles), en fonction du temps de parcous (en heures). La ligne horizontale grise correspond à 0.05, tandis que la ligne traitillée verticale indique la vraie médiane $\nu$ et la ligne pointillée verticale marque la vraie moyenne $\mu$.

puissance (Figure 1) pour les trois tests de localisation afin de répondre aux questions suivantes:

(a) Expliquez pourquoi la proportion de rejet de chaque test augmente quand on se déplace vers la droite sur le graphique.

(b) Supposez que l'on répète l'expérience de simulation, mais cette fois avec des sous-échantillons aléatoires de taille $n = 1000$. Comment est-ce que les points pour le test-$t$ pour un échantillon se compareraient à ceux tracés sur le graphique? Seraient-ils en dessous, à la même hauteur ou au dessus?

(c) Expliquez pourquoi la valeur sur le graphique pour le test-$t$ pour un échantillon **devrait être** approximativement 0.05 dans un voisinage de $\mu = 2.845$ (idem pour le test des signes et le test de Wilcoxon, où les valeurs devraient être approximativement 0.05 autour de $\nu = 2.833$).

(d) Selon la Figure 1, à quelle fréquence rejeteriez-vous l'hypothèse nulle pour le test des rangs signés de Wilcoxon à $\nu = 2.833$? Formulez une explication pour cette valeur inattendue sur la base de l'histogramme du temps de parcours et expliquez les conséquences de cette trouvaille sur votre inférence.

(e) Est-ce que le postulat du test-$t$ pour un échantillon est valide dans cet exemple? Produisez un diagramme quantile-quantile et commentez sur la robustesse du test-$t$ à des déviations de l'hypothèse de normalité.

4. Supposez que l'on veut comparer le tarif moyen pour les trains à grande vitesse pour les deux destinations, soit de Madrid vers Barcelone et le trajet inverse de Barcelone à Madrid. Une étude de simulation a été réalisée dans laquelle le test de Welch pour deux échantillons a été calculé sur des sous-échantillons aléatoires de taille $n = 1000$. Les données `renfe_simu` contiennent les différences moyennes (`difmoy`), les statistiques de test (`Wstat`), les valeurs-$p$ (`pval`) et les intervalles de confiance à 95% (`icbi` et `icbs`) pour 1000 répétitions. Supposez que l'on sait que la vraie différence moyenne dans la population est de $-0.28$€. Utilisez les données simulées pour répondre aux questions suivantes et **commentez brièvement** sur chaque sous-question.



Compiled 30/01/2020 at 10:12

   (a) Quel est le taux de couverture empirique des intervalles de confiance à 95% (c'est-à-dire le pourcentage des intervalles couvrant la valeur de la « vraie » différence moyenne)?
   (b) Tracez un histogramme des différences moyennes et superposez la vraie différence moyenne à l'aide d'un trait vertical.
   (c) Calculez la puissance du test (pourcentage de rejet de l'hypothèse nulle sous l'hypothèse alternative).
5. À l'aide des données `renfe`, testez si le prix moyen du billet pour un train de classe `AVE-TGV` est le même que celui d'un train régio-express (`REXPRESS`). Veillez à
   - énoncer l'hypothèse nulle et l'hypothèse alternative,
   - justifier avec soin le choix de votre statistique de test,
   - rapporter la différence moyenne estimée et un intervalle à 90% pour cette différence,
   - conclure dans le cadre de la mise en situation.
6. À l'aide des données `renfe`, comparez le prix des trains entrants (de Madrid à Barcelone) et des trains sortants (de Barcelone à Madrid) pour les trains à grande vitesse: en moyenne, est-ce qu'une direction est plus chère que l'autre? Comparez également le prix médian à l'aide d'un test nonparamétrique.
7. À l'aide des données `renfe`, testez si les billets pour les trains `AVE-TGV` sont plus chers les fins de semaine que les jours de semaine.
8. Ajustez un modèle linéaire pour expliquer le `prix` des billets `Promo` pour les trains à grande vitesse en fonction de `dest`, `classe`, `duree` et d'une variable additionelle indiquant si le jour de départ est une fin de semaine ou pas.
   (a) Écrivez l'équation du modèle théorique postulé.
   (b) Donnez les estimés des paramètres $\boldsymbol{\beta}$ et interprétez-les.
   (c) Testez la significativité de chaque variable et la significativité globale du modèle.
   (d) Produisez des diagnostics graphiques des résidus et commentez sur la validité des postulats du modèle linéaire (votre rapport devrait minimalement inclure un graphique des résidus et valeurs ajustées et un diagramme quantile-quantile des résidus studentisés externes).



Compiled 30/01/2020 at 10:12