

TGV espagnols: 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).

1.1 On considère le temps de parcours pour les trains à grande vitesse (AVE et AVE-TGV). Le temps médian entre les deux villes dans la « population » est de $v = 2.833$ 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. L'algorithme suivant a été répété 10 000 fois:

- sélection d'un sous-échantillon de taille $n = 100$.
- calcul de la statistique du test- t pour un échantillon correspondant à $\mathcal{H}_0 : \mu = \mu_0$ (versus $\mathcal{H}_0 : \mu \neq \mu_0$) pour différentes valeurs de μ_0 .
- calcul de la statistique du test des signes pour le test bilatéral $\mathcal{H}_0 : v = v_0$ pour différentes valeurs de v_0 .
- calcul de la statistique du test des rangs signés de Wilcoxon pour le test bilatéral $\mathcal{H}_0 : v = v_0$ pour différentes valeurs de v_0 .
- 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 $\mathcal{H}_0 : \mu = \mu_0$ contre l'alternative bilatérale à $\mu_0 \in \{2,83; \mu; 2,835; 2,84; \dots; 2,995; 3\}$ (pour le test des signes et de Wilcoxon, nous testons si la médiane est égale à ces même valeurs). Utilisez la courbe de puissance (Figure 1) pour les trois tests de localisation afin de répondre aux questions suivantes:

- Expliquez pourquoi la proportion de rejet de chaque test augmente quand on se déplace vers la droite sur le graphique.
- 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?
- 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 $v = 2,833$).

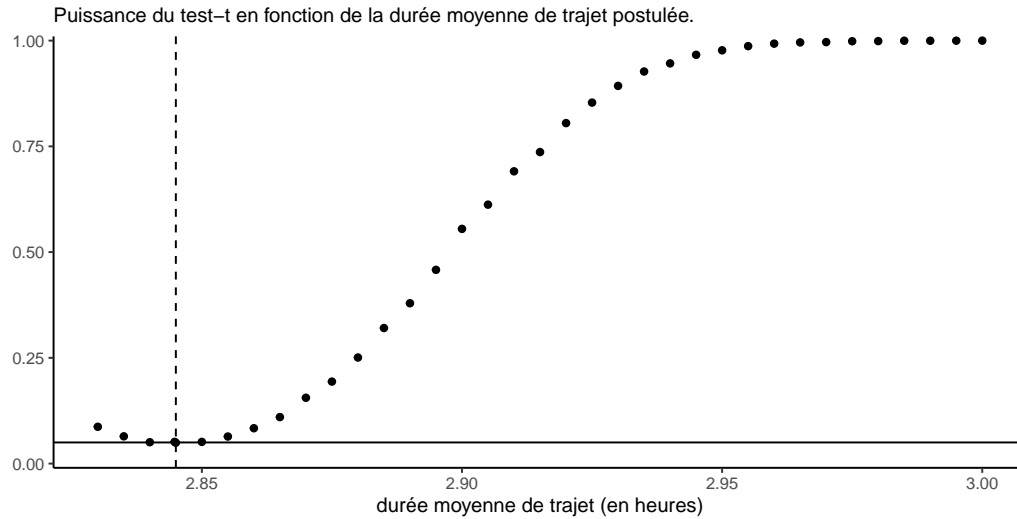


Figure 1: Courbe de puissance pour trois tests de localisation, soit le test- t pour un échantillon (disque), le test des rangs signés de Wilcoxon (croix) et le test des signes (cercles), en fonction du temps de parcours (en heures). La ligne horizontale grise correspond à 0,05, tandis que la ligne traitillée verticale indique la vraie médiane v et la ligne pointillée verticale marque la vraie moyenne μ .

- (d) Selon la Figure 1, à quelle fréquence rejeteriez-vous l'hypothèse nulle pour le test des rangs signés de Wilcoxon à $v = 2,833$? Expliquez les conséquences de cette trouvaille sur votre inférence.
- (e) Produisez un diagramme quantile-quantile normal et commentez sur la robustesse du test- t à des déviations de l'hypothèse de normalité.
- 1.2 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 (`valp`) 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 $-0,28\text{€}$. Utilisez les données simulées pour répondre aux questions suivantes et **commentez brièvement** sur chaque sous-question.
- 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)?
 - Tracez un histogramme des différences moyennes et superposez la vraie différence moyenne à l'aide d'un trait vertical.
 - Calculez la puissance du test (pourcentage de rejet de l'hypothèse nulle sous l'hypothèse alternative).
- 1.3 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 régional-express (REXPRESS). Veuillez 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.