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 $\nu = 2.833$ 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. L'algorithme suivant a été répété 10 000 fois:

(a) sélection d'un sous-échantillon de taille $n = 100$.

(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 μ_0.

(c) calcul de la statistique du test des signes pour le test bilatéral $\mathscr{H}_0 : \nu = \nu_0$ pour différentes valeurs de ν_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 ν_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 $\mu_0 \in \{2,83; \mu; 2,835; 2,84; \ldots; 2,995; 3\}$ (pour le test des signes et de Wilcoxon, nous testons si la médiane est égale à ces même valeurs). Utilisez la courbe de 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 $\nu = 2,833$).

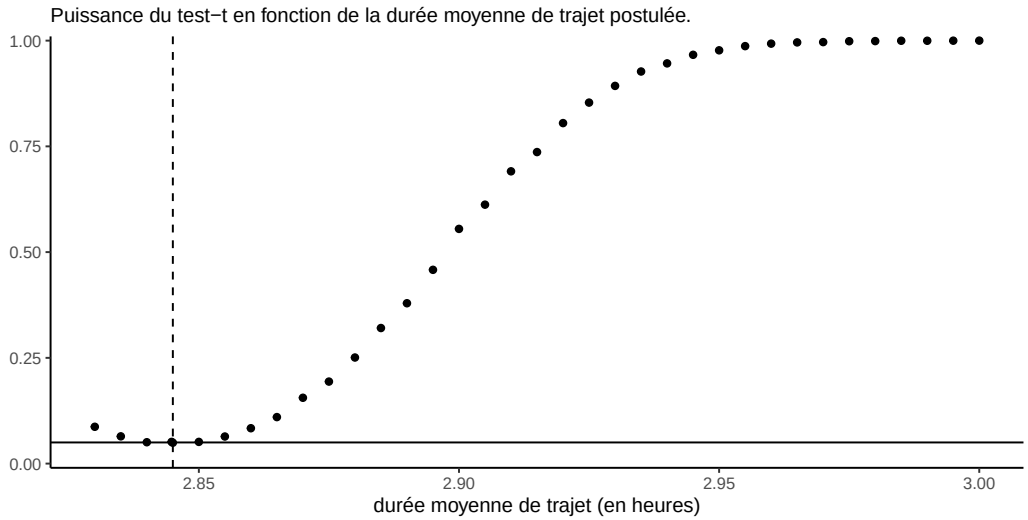

Figure 1: Courbe de puissance pour trois tests de localisation, soit le test-t pour un échantillon (disque), le test des rangs signés de Wilcoxon (croix) et le test des signes (cercles), en fonction du temps de parcours (en heures). La ligne horizontale grise correspond à 0,05, tandis que la ligne traitillée verticale indique la vraie médiane ν 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 $\nu = 2,833$? 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.

Solution

(a) La courbe donne le pourcentage de rejet de l'hypothèse nulle pour le test-t pour un échantillon. Le plus on s'éloigne de la vraie valeur μ, le plus de preuves on accumule pour détecter un départ de l'hypothèse nulle $\mathscr{H}_0$. Puisque le test est fait à niveau $\alpha = 0,05$, la courbe tend vers 0,05 près de μ et augmente vers un des deux côtés à mesure que l'on s'éloigne de la vraie moyenne.

(b) La puissance augmente si n croît, donc on s'attend que la courbe soit au dessus partout, sauf dans un voisinage de μ où elle devrait être approximativement 0,05 si les hypothèses du test sont respectées.

(c) Les données ne sont pas normalement distribuées et fortement discrétisées, mais la courbe de puissance du test-t pour un échantillon semble augmenter à mesure que l'on s'éloigne de μ et le niveau nominal du test correspond au niveau empirique, ou erreur de type I. Cela illustre la robustesse du test au départ de la normalité et c'est une conséquence du théorème central limite.

(d) Le niveau du test α, ici 5%, représente le pourcentage de rejet de l'hypothèse nulle si cette dernière est vraie.

(e) L'erreur de type I du test des rangs signés de Wilcoxon est 0,44, très loin du niveau nominal de 0,05. La puissance augmente à mesure qu'on s'éloigne de la vraie médiane ν, mais l'absence de symmétrie et les duplicatas bousille les propriétés du test quand $\mathscr{H}_0 : \nu = \nu_0$, ce qui donne une erreur de Type I enflée; cela démontre que les tests nonparamétriques ne sont pas une panacée.

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

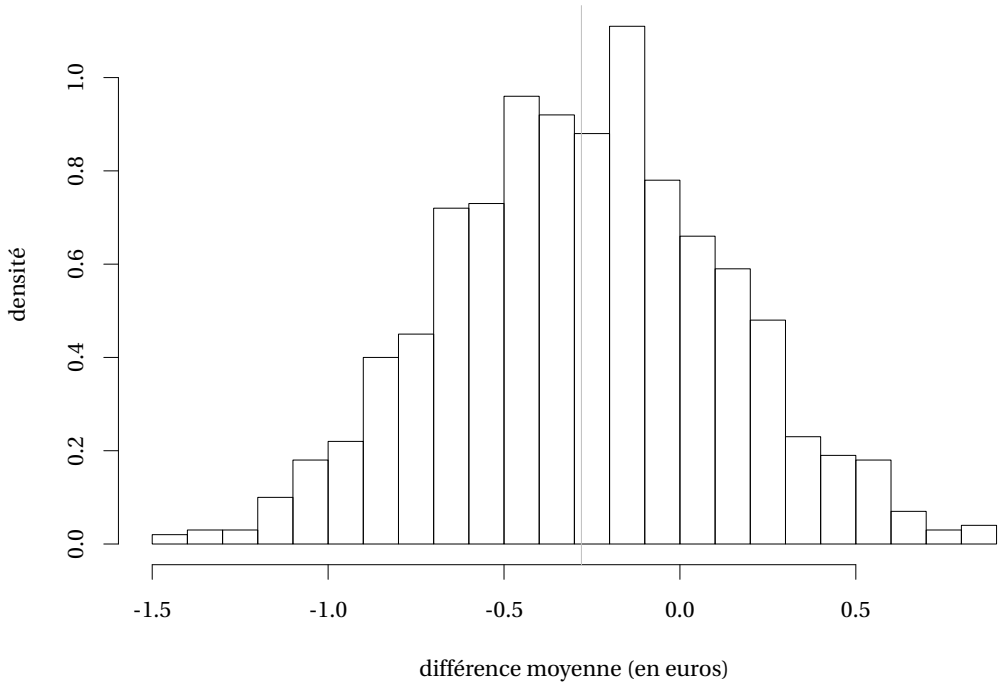

Figure 2: Histogramme de la différence de prix moyenne pour les trains à grande vitesse Madrid–Barcelone versus Barcelone–Madrid avec la moyenne de l'échantillon (trait vertical gris).

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 $-0,28$€. Utilisez les données simulées pour répondre aux questions suivantes et **commentez brièvement** sur chaque sous-question.

(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).

Solution

(a) Le taux de couverture empirique est 0,947; cette valeur est près du taux de couverture théorique nominal, ce qui indique que le test est bien calibré.

(b) Figure 2 contient deux histogrammes: la différence moyenne semble approximativement normale et centrée en 0.28, tandis que les valeurs-p sont réparties dans l'intervalle $[0,1]$ avec plus de valeurs près de zéro.

(c) La puissance est $0,105$. Sous le régime alternatif (puisque $\Delta = 0.28$ €), on rejette 10,5% du temps l'hypothèse nulle. Ce pourcentage est faible parce que la différence est petite et donc difficile de distinguer cette différence de la variabilité intrinsèque de la statistique à moins d'avoir une grande taille d'échantillon. La différence moyenne estimée avec l'échantillon est de $0,274$.

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 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.

Solution
Le prix des billets `REXPRESS` est fixe et vaut 43.25€, on a donc un échantillon aléatoire que pour l'autre classe de train!

- L'hypothèse nulle est $\mathscr{H}_0 : \mu_{\texttt{AVE-TGV}} = 43,25$€contre l'alternative $\mathscr{H}_1 : \mu_{\texttt{AVE-TGV}} \neq 43,25$€, où $\mu_{\texttt{AVE-TGV}}$ est le prix moyen d'un billet de train `AVE-TGV`.
- Puisqu'on veut comparer la moyenne et qu'un seul échantillon est aléatoire, on utilise un test-t pour un échantillon.
- La différence moyenne estimée est $45,63$€$= 88,88$€$-43,25$€, avec un intervalle de confiance à 90% pour la différence moyenne de $[44,14; 47,12]$.
- La statistique t, qui vaut 50,519 ici, suit une loi Student-t avec 428 degrés de liberté; une approximation normale serait identique. La valeur-p associée est négligeable, on conclut que le prix des trains `AVE-TGV` et `Rexpress` diffèrent.

Compilé le 29/08/2024 à 09:08