1.1 **Prix de billets de 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`).

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 $\nu = 2,833$).
(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 rejection 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

Compilé le 14/12/2020 à 15:50

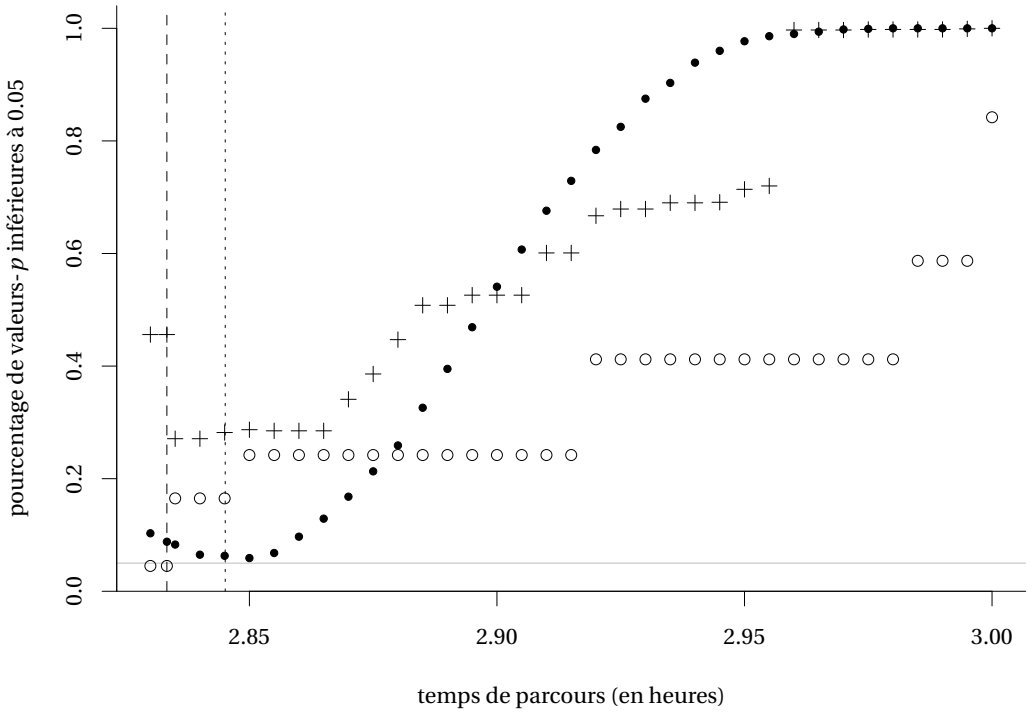

Figure 1: Courbe de puissance pour trois tests de localisation, soit le test-t pour un échantillon (disque), le test des rangs signés de Wilcoxon (croix) et le test des signes (cercles), en fonction du temps de parcours (en heures). La ligne horizontale grise correspond à 0,05, tandis que la ligne traitillée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

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. 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) Le niveau empirique 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 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.

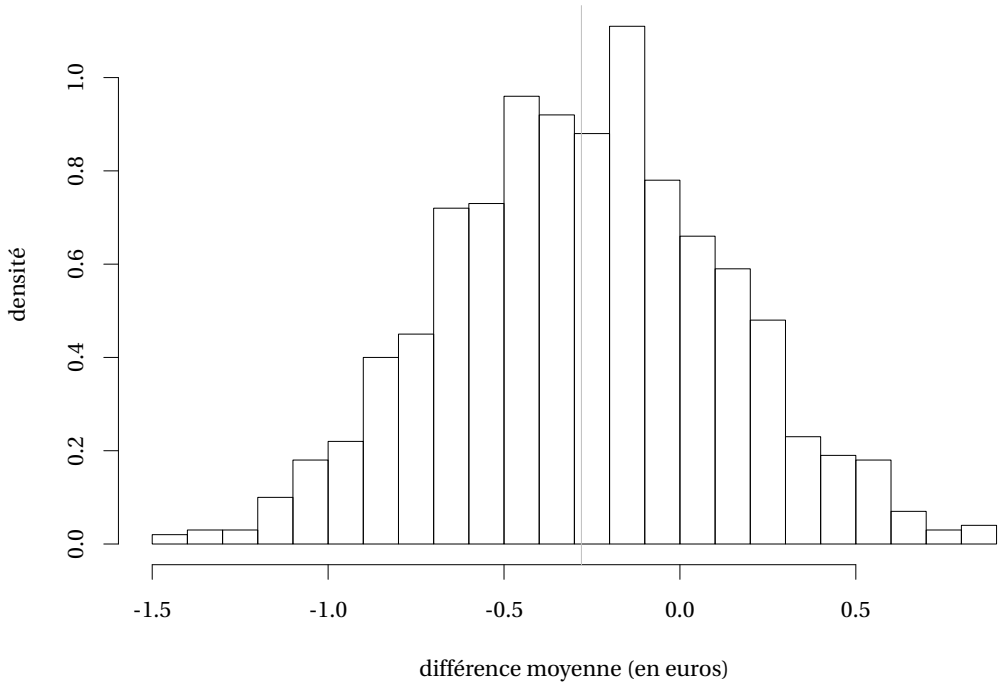

Figure 2: Histogramme de la différence de prix moyenne pour les trains à grande vitesse Madrid–Barcelone versus Barcelone–Madrid avec la moyenne de l'échantillon (trait vertical gris).

- 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.

1.4 Les données fictives `assurance` font partie du livre "Machine Learning with *R*" de Brett Lantz (2003). La base de donnée contient des informations sur les frais médicaux facturés à 1338 adultes américains assurés au courant de l'année 2003, soit

- `age`: âge (en années)
- `sexe`: sexe, `homme` ou `femme`,
- `imc`: indice de masse corporelle (en kg/m^2),
- `enfant`: nombre d'enfants à charge,
- `fumeur`: `oui` pour les fumeurs, `non` autrement,
- `region`: lieu de résidence, une région parmi `sudouest`, `sudest`, `nordouest` ou `nordest`,
- `frais`: les frais médicaux annuels (en dollars USD).

Selon l'Organisation mondiale de la Santé (OMS), l'indice de masse corporelle (IMC) permet de classer les individus conformément à une échelle allant de l'insuffisance pondérale à l'obésité morbide (classe III). Ladite classification est définie dans le tableau Table 1.

À l'aide des données `assurance`, répondez aux questions suivantes.

(a) Effectuez une analyse exploratoire des données: quelles sont les aspects les plus importants pour expliquer les frais médicaux?

Solution

i. La distribution des `frais` est strictement positive, asymmétrique à droite (la moyenne est plus grande

Classification	IMC (kg/m^2)
< 18.5	Insuffisance pondérale
18.5–24.9	Corpulence normale
25.0–29.9	Surpoids
30.0–34.9	Obésité
35.0–39.9	Obésité de classe II et III

Table 1: Classification internationale de l'obésité chez les adultes selon l'OMS.

que la médiane). On note la présence de certaines valeurs aberrantes.

ii. Les graphiques présentes dans la Figure 3 montrent une augmentation linéaire des frais avec l'âge, mais il y a apparence de trois groupes avec une plus forte hétérogénéité pour ceux qui ont des primes plus élevée. Tous les individus qui sont dans le groupe avec les frais les plus élevés sont fumeurs. L'indice de masse corporel semble expliquer uniquement les frais une fois combiné au statut de fumeur: la surprime survient seulement pour les personnes obèses, soit celles dont l'IMC est égal ou supérieur à 30kg/m^2 selon la définition de l'OMS.

(b) Les fumeurs paient-ils des frais médicaux en moyenne équivalents aux non-fumeurs? Justifiez adéquatement votre réponse

Solution

Non, les fumeurs paient en moyenne des frais significativement plus élevés tel qu'illustré par la Figure 3 qui montrent une forte hétérogénéité et des variances inégales; le test de Levene pour l'égalité des variances confirment cette affirmation, avec un intervalle de confiance à 95% pour le rapport des variances de [0,22; 0,32]. On rejete l'hypothèse nulle que $\mu_F = \mu_N$ en faveur de l'alternative $\mu_F \neq \mu_N$, où μ_F (μ_N) représente la moyenne des frais médicaux (en dollars américains) pour les fumeurs (resp. non-fumeurs). La statistique de Welch est $32,75$ et la valeur-p est plus petite que 10^{-15}; même si les données ne sont pas normales, les conclusions sont sans équivoque à cause de la taille de l'échantillon et de la différence moyenne estimée de 23616$.

(c) Les fumeurs considérés obèses (`imc` ≥ 30) paient-ils des frais médicaux en moyenne plus élevés que les fumeurs non obèses? Donnez trois intervalles de confiance à 90%, 95% et 99% pour estimer la différence moyenne dans ce contexte. Comparez les intervalles et expliquez les différences observées selon le niveau.

Solution

Oui, les fumeurs obèses paient des frais plus élevés que les fumeurs non-obèses. Le panneau droit en bas de la Figure 3 montre clairement ce fait, mais un test formel de l'hypothèse unidirectionnelle $\mathscr{H}_0 : \mu_0 \leq \mu_1$ contre l'alternative $\mathscr{H}_1 : \mu_0 > \mu_1$, où μ_0 (μ_1) sont les frais moyens pour un fumeur obèse (non-obèse). Les intervalles de confiance dérivés sur la base de la statistique de Welch à niveau 90%, 95% et 99% sont respectivement $(-\infty; -19333,08)$, $(-\infty; -19087,71)$ et $(-\infty; -18625,11)$. Plus le niveau du test α est grand, plus la borne supérieure est large et plus les intervalles sont courts; à noter que ces derniers sont également imbriqués.

1.5 **Calcul de la puissance d'un test statistique**

Le programme SAS `puissance.sas`, écrit par Rick Wicklin de SAS Institute Inc., contient du code pour faire une étude de simulation afin de calculer la puissance d'un test-t pour deux échantillons (l'équivalent d'un modèle linéaire simple avec une variable binaire comme variable explicative).

(a) Expliquez brièvement dans vos mots en quoi consiste les étapes de l'étude de simulation.

Solution

L'étude Monte Carlo sert à calculer la puissance d'un test-t pour deux échantillons $\boldsymbol{Y}$ et $\boldsymbol{Z}$ en fonction de leur différence de moyenne. Plus spécifiquement, on s'intéresse à $\mathscr{H}_0 : \Delta = 0$ contre l'alternative bilatérale

pour différentes valeurs de Δ. Pour chaque valeur de Δ, les données des deux échantillons de taille $n = 10$ sont indépendantes et simulées de lois normales, $Y_i \sim \mathsf{No}(\mu, 1)$ et $Z_i \sim \mathsf{No}(\mu+\Delta, 1)$. Pour chacun des B échantillons, on simule des données, on calcule la statistique de test et on calcule une valeur binaire qui vaut un si on rejette à niveau 5% l'hypothèse nulle, selon la loi de référence $\mathsf{St}(n-2)$. On répète ce calcul B fois pour chaque Δ et on retourne un estimé de la puissance qui correspond à la moyenne empirique des données binaires, soit la proportion de rejet de $\mathscr{H}_0$.

(b) Tracez le graphique pour les paramètres $n_1 = n_2 = 10$, $\sigma = 1$ et $B = 10\,000$ et commentez sur l'apparence de ce dernier.

Solution

La courbe en V est symmétrique autour de $\Delta = 0$. Si $\Delta = 0$, l'hypothèse nulle est vraie: on devrait obtenir le niveau du test comme proportion de rejet puisque les données sont normales et homoscédastiques, ici 5% par défaut. À mesure que $|\Delta|$ augmente, la puissance augmente et la puissance atteint approximativement 1 quand $|\Delta| > 2$.

(c) Faites varier le nombre de simulations de $B = 100$ à $B = 10\,000$. Que remarquez-vous quand le nombre de simulation est petit? Expliquez pourquoi cet effet disparaît quand la nombre de simulations augmente.

Solution

On approxime la vraie proportion de rejet π_i pour chaque valeur Δ_i à l'aide d'un échantillon d'une loi binomiale avec B essais: l'estimateur de la moyenne est sans biais pour π_i, la probabilité de rejet quand $\Delta = \Delta_i$. Avec $B = 100$, il y a beaucoup de variabilité (la variance de la moyenne est $\pi_i(1-\pi_i)/B$) et donc la courbe de la puissance avec $B = 100$ est discontinue et fluctue avec Δ, plutôt que d'augmenter de façon monotone croissante comme lorsque $B = 10\,000$.

(d) Modifiez le code pour que la taille des groupes soit $n_1 = 10, n_2 = 30$ et $n_1 = 20, n_2 = 20$. Dans lequel des deux scénarios la puissance est-elle plus élevée et pourquoi?

Solution

La puissance est plus élevée quand les tailles d'échantillons sont égales (balancées). Les estimateurs de la moyenne et de la variance conjoint sont sans biais, peu importe la taille de l'échantillon, parce que les données sont simulées de lois normales et sont homoscédastiques. La raison pour l'augmentation de la puissance dans le cas $n_1 = n_2 = 20$ plutôt que $n_1 = 10$, $n_2 = 30$ ne vient donc pas du nombre de données (égal dans les deux cas à $n = 40$), mais plutôt au fait que l'estimateur conjoint de la variance est

$$S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

On remarque que $1/10 + 1/30 = 4/30 > 3/30 = 1/20 + 1/20$, d'où la puissance plus élevée dans le second cas (variance plus faible, statistique plus élevée pour le même ensemble de paramètres). Règle générale, cette affirmation n'est vraie que si la variance des deux échantillons est égale.

(e) Modifiez le code pour simuler $n = 20$ observations dans chaque groupe de lois normales d'écart-type $\sigma_1 = 1$ et $\sigma_2 = 5$ avec $B = 100\,000$. Rapportez l'estimé de l'erreur de type I avec le nombre de simulations et un intervalle de confiance ponctuel à 90% pour l'erreur de type I; expliquez comment vous avez dérivé ce dernier.

Solution

Puisque chaque réplication est indépendante et la probabilité de rejet la même (même conditions expérimentales), le nombre de succès suit une loi binomiale avec B essais et probabilité de succès (inconnue) p. Comme il s'agit d'une étude Monte Carlo, le nombre de succès est aléatoire. Avec 100 000 réplications, l'estimé est de $\widehat{p} = 0.05512$, loin de la valeur postulée du test à niveau 5%. Un intervalle de confiance à 90% de Wald est

$$\widehat{p} \pm \Phi^{-1}(0.95)\sqrt{\widehat{p}\cdot(1-\widehat{p})/B} = 0.05512 \pm 1.64485\sqrt{0.05512\cdot 0,94488/100000} = [0.05280.0574]$$

Cela m'amène à conclure qu'à niveau 90%, l'erreur de type I est significativement différente de la valeur postulée du niveau (le test suppose que les variances sont identiques, donc les postulats de validité ne sont pas valide et l'erreur de type I est légèrement différente du niveau du test).
On pourrait aussi ajuste un modèle linéaire généralisé avec une loi binomiale. L'intervalle de confiance basé sur le rapport de vraisemblance est alors $[0.0537; 0.0565]$.

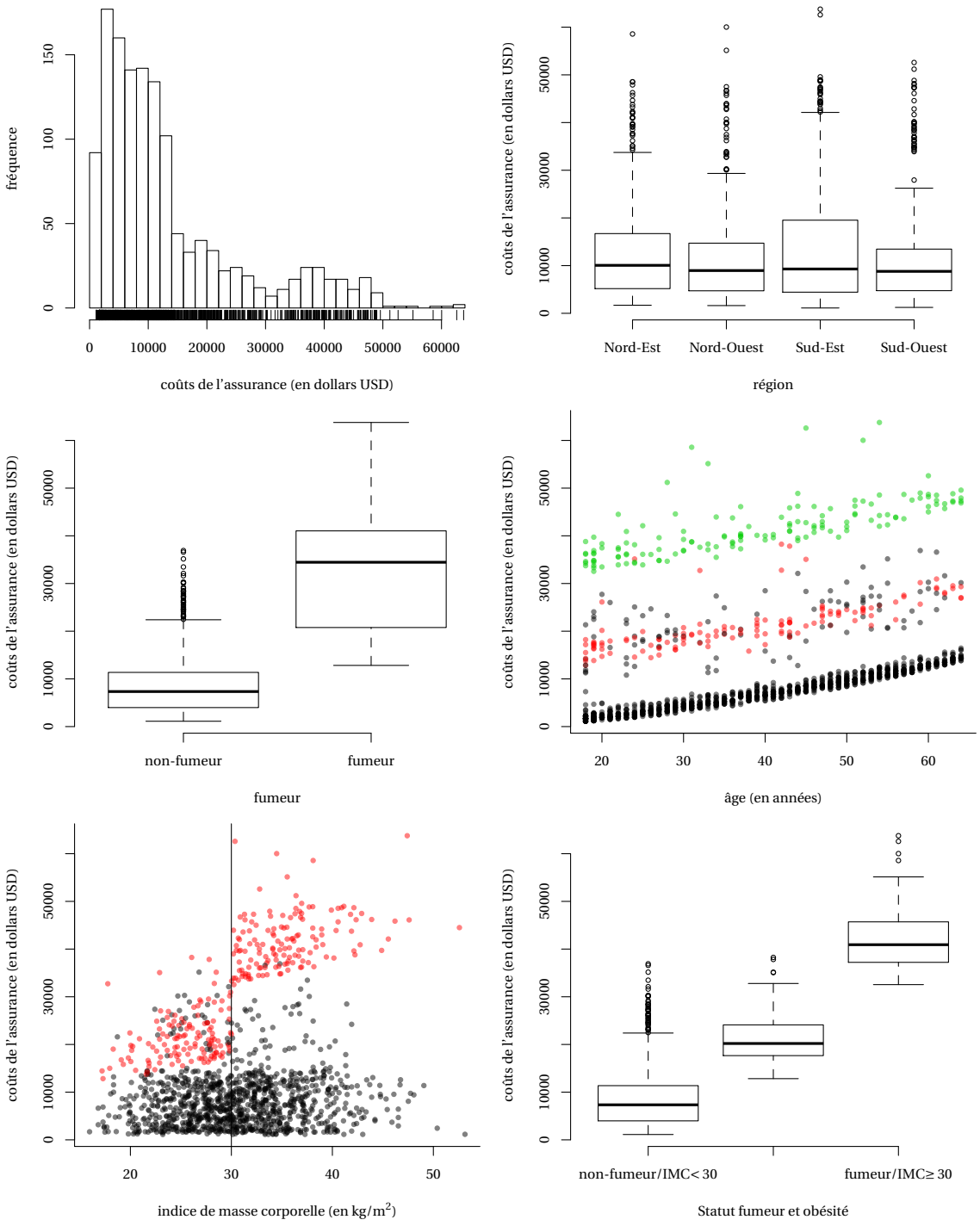

Figure 3: Histogramme des frais médicaux annuels (en dollars américains) (haut, gauche), boîte à moustache des frais en fonction du sexe (haut, gauche) et par statut de fumeur/non-fumeur (milieu, gauche). Nuages de points des frais contre l'âge (milieur, droit), qui montre une traîne linéaire pour trois groupes correspondants aux non-fumeurs et non-obèses (noir), fumeurs non-obèses (rouge) et fumeurs obèses (vert). Boîte à moustache par fumeur (bas, droite). Nuage de point des frais en fonction de l'indice de masse corporelle avec fumeurs (rouge) et non-fumeurs (noir); la ligne verticale indique le seuil d'obésité (bas, gauche).