2.1 **Test-t unidirectionnel** Un avantage des tests unidirectionnels est que le test est plus puissant pour détecter l'hypothèse spécifiée — c'est-à-dire que la probabilité de rejeter $\mathscr{H}_0$ sera plus grande si $\mathscr{H}_0$ est fausse. Pour faire un test unidirectionnel avec la procédure `ttest` de SAS , on utilise l'option `sides=l` (pour "lower").

Il est important de comprendre que la spécification des hypothèses (unidirectionnelles ou bidirectionnelles) se fait habituellement au début du processus de recherche, avant même la collecte des données. Par conséquent, il n'est pas justifié de regarder les données, et de voir quelle contre-hypothèse tester, car en faisant cela, le vrai niveau du test et la valeur-p sont faussés.

Dans le cas de l'exemple du cours sur le paiement par carte de crédit versus comptant, supposons que nous voulons tester l'hypothèse que les gens paient le mme montant peu importe le type de paiement.

(a) Écrivez les hypothèses $\mathscr{H}_0$ et $\mathscr{H}_1$ du modèle.

Solution

L'hypothèse nulle est $\mathscr{H}_0 : \mu_A = \mu_B$ et la contre-hypothèse $\mathscr{H}_1 : \mu_A \neq \mu_B$, où $\mu_A(\mu_B)$ est le montant moyen que les consommateurs sont prêts à dépenser pour une paire de billet payée comptant (par carte de crédit).

(b) Testez cette hypothèse à l'aide de SAS et concluez.

Solution

La valeur-p du test bilatéral de Welch pour deux échantillons est 0.0017, soit deux fois la valeur-p du test unilatéral (à cause de la symétrie). On rejette l'hypothèse nulle, donc le montant moyen offert est différent à niveau 5% pour le groupe qui paie par carte de crédit et celui qui paie comptant. L'intervalle de confiance bilatéral à 95% est $(-23.99, -6.02)$, tandis que l'intervalle unilatéral est $(-\infty, -7.52)$; on s'attend à ce que la différence moyenne soit dans un tel intervalle 95% du temps en échantillons répétés.

2.2 **Influence des valeurs extrêmes dans le test-t** Le test-t est sensible à la présence de valeurs extrêmes dans les données tandis que le test de la somme des rangs de Wilcoxon est plus robuste. Le but de cet exercice est d'illustrer la sensibilité du test-t aux valeurs aberrantes en reprenant l'exemple du paiement par carte de crédit versus comptant.

Ici, nous allons artificiellement remplacer la première observation de l'ensemble de données, qui vaut 62$, par la valeur 210$. Ceci peut se faire à l'aide du code SAS suivant:

```
data temp;
set infe.billets;
if _N_=1 then offre=210;
run;
```

En pratique, les valeurs aberrantes peuvent être détectées et éliminées lors d'une analyse exploratoire (par exemple, les valeurs manquantes sont souvent codées avec -1 ou 999 même si ces valeurs sont impossibles). Si la valeur observée est extrême, mais plausible, il est difficile de justifier son retrait.

(a) Effectuez un test-t pour deux échantillons avant et après le changement. Commentez sur la différence des résultats au niveau des intervalles de confiance pour la différence des moyennes, au niveau des valeurs-p du test et au niveau des résultats du test d'égalité des variances.

Solution

Dans les données originales, la variance des montants offerts pour les billets par crédit et celle des offres en argent comptant sont significativement différentes (valeur-p de moins de 10^{-4}). Avec la valeur aberrante, la valeur-p du test d'égalité des variances est 0.369 (mais l'ajustement est mauvais, voir le diagramme quantile-quantile). Avec la valeur aberrante, la valeur-p du test d'égalité des moyennes (test de Welch) passe de 0.0009, à 0.0532 pour les données modifiées.

(b) Identifiez la présence de la valeur extrême sur les boîtes-à-moustache et histogrammes des données fournis par la procédure `ttest`.

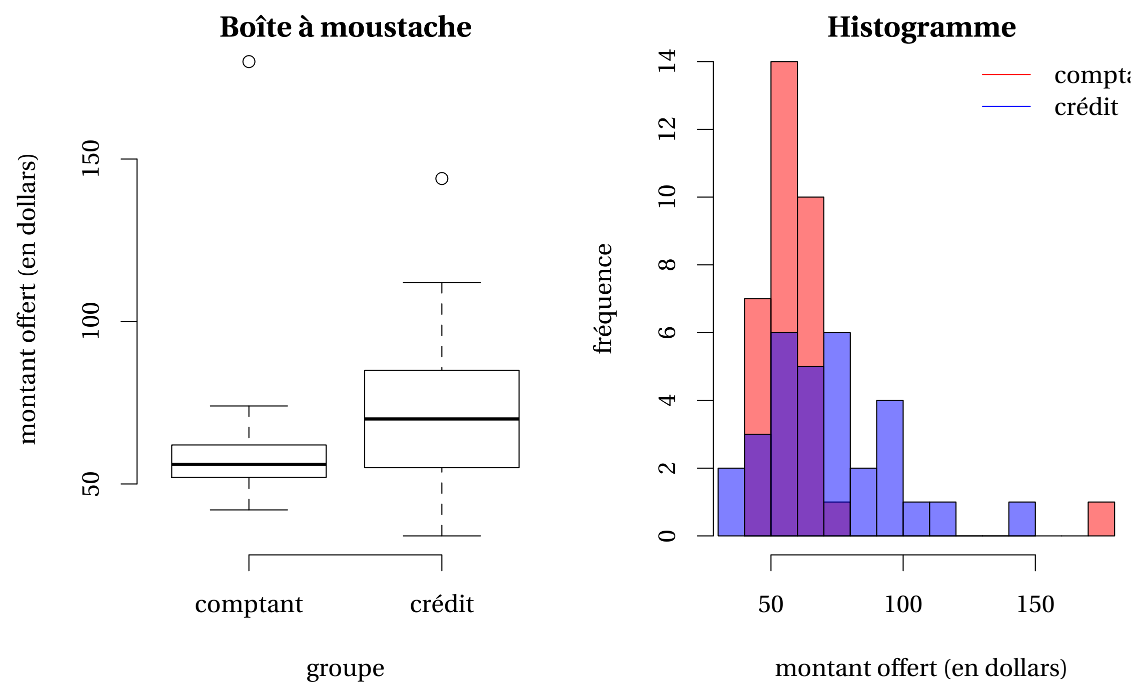

Figure 1: Boîte à moustache (gauche) et histogramme (droit) pour les données `billets` modifiées.

Solution

La valeur aberrante est clairement visible Figure 1.

(c) Refaites la même analyse avant et après le changement dans les données à l'aide du test de la somme des rangs de Wilcoxon. Est-ce que les conclusions de votre test changent?

Solution

Puisqu'il y a des ex aequo dans les données, la valeur-p n'est pas unique, mais est elle approximativement 0.0014 et augmente à 0.0032 avec la valeur aberrante, sans affecter la conclusion du test.

2.3 Les données de cet exercice sont inspirées de l'article

Zellner *et al.* (2010). Art on the Plate: Effect of Balance and Color on Attractiveness of, Willingness to Try and Liking for Food, *Food Quality and Preference*, 21(**5**), 575–578.

Cet article traite du lien entre l'aspect visuel d'un plat (symétrie et couleur) sur l'intention de goûter, l'attraction et le plaisir procuré. Le fichier de données `nourriture.sas7bdat` contient des données simulées propre à cette étude. Il contient 68 sujets et les cinq variables suivantes :

- `balance`: symétrie du plat, soit symétrique (1) ou non symétrique (2);
- `couleur`: couleur du plat, soit monochrome (1), soit coloré (2);
- `attraction`: score d'attraction entre −100 et 100 (les valeurs négatives indiquant une répulsion, les valeurs positives indiquant une attraction)
- `desir`: score relié au désir de goûter le plat entre −100 et 100 (les valeurs négatives indiquant une aversion, les valeurs positives indiquant un désir)
- `plaisir`: score relié au plaisir que le sujet a eu à goûter le plat entre −100 et 100 (les valeurs négatives indiquant un désagrément, les valeurs positives indiquant un plaisir).

Vous allez comparer le score de la variable `desir` entre les deux niveaux de la variable `couleur`. Pour ce faire,

(a) Calculez les moyennes et les variance de `desir` dans les deux groupes (monochrome et coloré). Qu'observez-vous?

Solution
Les moyennes semblent différentes, mais leur variabilité est trop large pour conclure quoi que ce soit sur la base d'une simple observation. Les scores moyens (erreurs-types) sont 18.06 (5.37) pour les plats monochrome et 30.53 (4.57) pour les plats colorés.

(b) Écrivez formellement l'hypothèse statistique que vous voulez tester.

Solution
L'hypothèse de recherche est $\mathscr{H}_0 : \mu_M = \mu_C$ contre l'alternative $\mathscr{H}_1 : \mu_M \neq \mu_C$, où $\mu_M(\mu_C)$ est le score moyen sur une échelle de $[-100, 100]$ pour l'assiette monochrome (colorée).

(c) Effectuez le test approprié, concluez et interpréter les résultats.

Solution
On ne rejette pas l'hypothèse nulle que les variances sont égales (valeur-p du test de Levene de 0.357). Que l'on utilise le test de Welch ou le test-t pour deux échantillons, la valeur-p est de 0.215 à trois décimales près. On ne rejette pas l'hypothèse nulle que l'attractivité moyenne ne dépend pas de la coloration du plat.

(d) Décrivez la différence dans le score entre les deux groupes à l'aide d'un estimateur de la différence moyenne ainsi que d'un intervalle de confiance.

Solution
La différence moyenne entre le score d'attraction des assiettes colorées et monochromes est -12.4706. L'intervalle de confiance à 95% pour la différence moyenne, obtenu en estimant la variance indépendamment dans chaque groupe, est $[-32.3901, 7.4489]$.

(e) Vérifiez graphiquement la normalité des données dans chaque groupe et concluez sur la validité de votre test.

Solution
Les diagrammes quantiles-quantiles dans chaque groupes n'indiquent aucune déviation apparente de normalité même si le groupe d'assiettes colorées a quelques petites valeurs plus grande qu'attendu.

(f) Reprendre les mêmes analyses que précédemment en utilisant cette fois le test de la somme des rangs de Wilcoxon ou le test de Fligner–Policello. Comparez vos résultats avec ceux du test-t classique.

Solution
Le test de la somme des rangs (valeur-p de 0.2108) et le test de Fligner–Policello (valeur-p de 0.2143) sont également applicables et donnent des résultats quasi-identiques.

2.4 Le fichier de données `Assurances` contient, entre autres, de l'information sur les frais médicaux facturés à 1338 adultes américains assurés au courant de l'année 2003. Les données simulées ont été extraites du livre "Machine Learning with *R*" de Brett Lantz (2003) et contiennent les informations suivantes

- `age`: âge (en années)
- `sexe`: sexe, `homme` ou `femme`,
- `imc`: indice de masse corporelle (en $\mathrm{kg/m^2}$),
- `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 2013 (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.

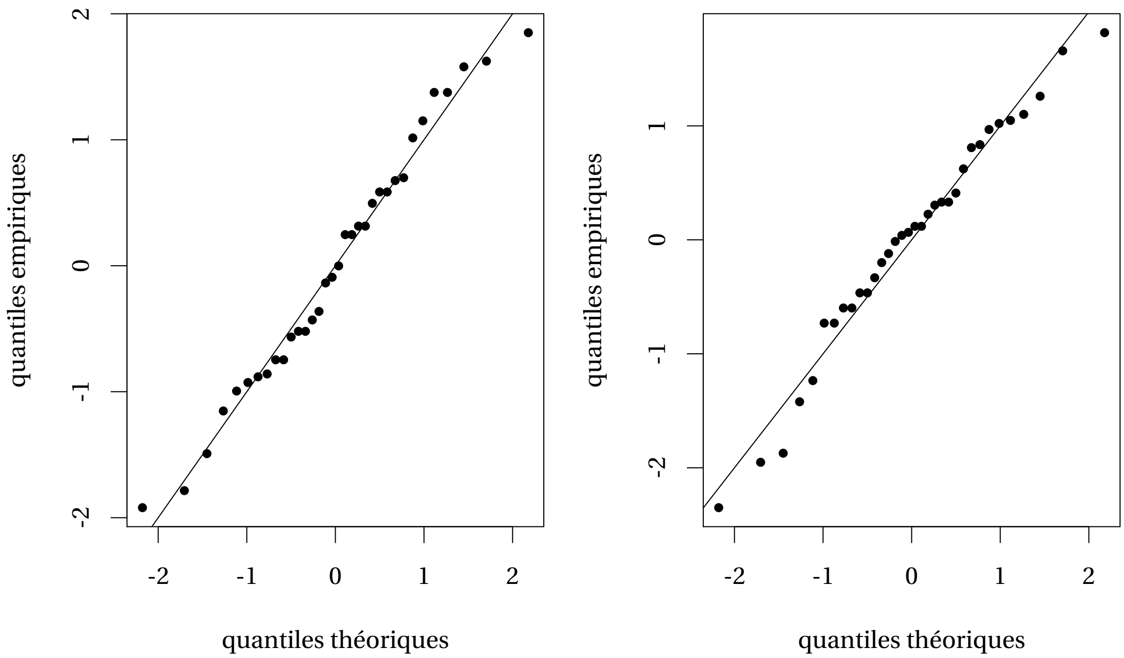

Figure 2: Diagramme quantile-quantile normal pour l'attractivité selon la couleur (coloré ou monochrome) pour les données `nourriture`.

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'OMC.

(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 que la médiane). On note la présence de certaines valeurs aberrantes. La distribution de `imc` est en revanche symmétrique en l'apparence et centrée autour de 30.

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.

(d) Existe-t-il une différence moyenne statistiquement significative entre l'indice de masse corporelle chez les hommes et les femmes? Un test non paramétrique est-il justifié dans ce contexte? Si oui, comparez les résultats avec un test-t à niveau $\alpha = 0.05$.

Solution

Dénotons l'indice de masse corporelle (en kg/m^2) médian par ν_h pour les hommes et ν_f pour les femmes, avec leur variance respective σ_h^2 et σ_f^2. Nous avons deux alternatives nonparamétriques, soit le test de la somme des rangs de Wilcoxon et le test de Fligner–Policello. L'égalité des distributions pour la validité du premier peut être vérifiée graphiquement dans la Figure 4 et on peut également tester l'égalité des variance — on ne rejette l'hypothèse nulle $\mathscr{H}_0 : \sigma_h^2 = \sigma_f^2$ contre la contre-hypothèse $\sigma_h^2 \neq \sigma_f^2$, ce qui nous console dans notre diagnostic graphique.

La valeur-p pour l'hypothèse nulle $\nu_h = \nu_f$ contre l'hypothèse alternative bilatérale $\nu_h \neq \nu_f$ est de $0,1012$ pour le test de Wilcoxon et de $0,1014$ pour le test de Fligner–Policello. On ne rejette pas l'hypothèse nulle à niveau 5%, et on conclut qu'il n'y pas de preuve de différence dans l'indice de masse corporel selon le sexe de l'individu. En comparaison, le test de Welch donne un valeur-p de 0.09.

(e) Y a-t-il une différence moyenne statistiquement significative entre l'indice de masse corporelle des habitants du Nord comparativement à ceux du Sud?

Solution

Soit μ_N et μ_S les indices de masse corporels moyens (en kg/m^2) pour les Américains du Nord et du Sud, respectivement. Les variances des deux groupes sont inégales et le test de Welch pour deux échantillons donne une valeur-p de 10^{-15}; l'intervalle de confiance symmétrique à 95% pour la différence est $(2,2;3,5)$. On rejette $\mathscr{H}_0 : \mu_N = \mu_S$ en faveur de l'alternative $\mu_N \neq \mu_S$ et donc il y a une différence significative d'IMC selon la région.

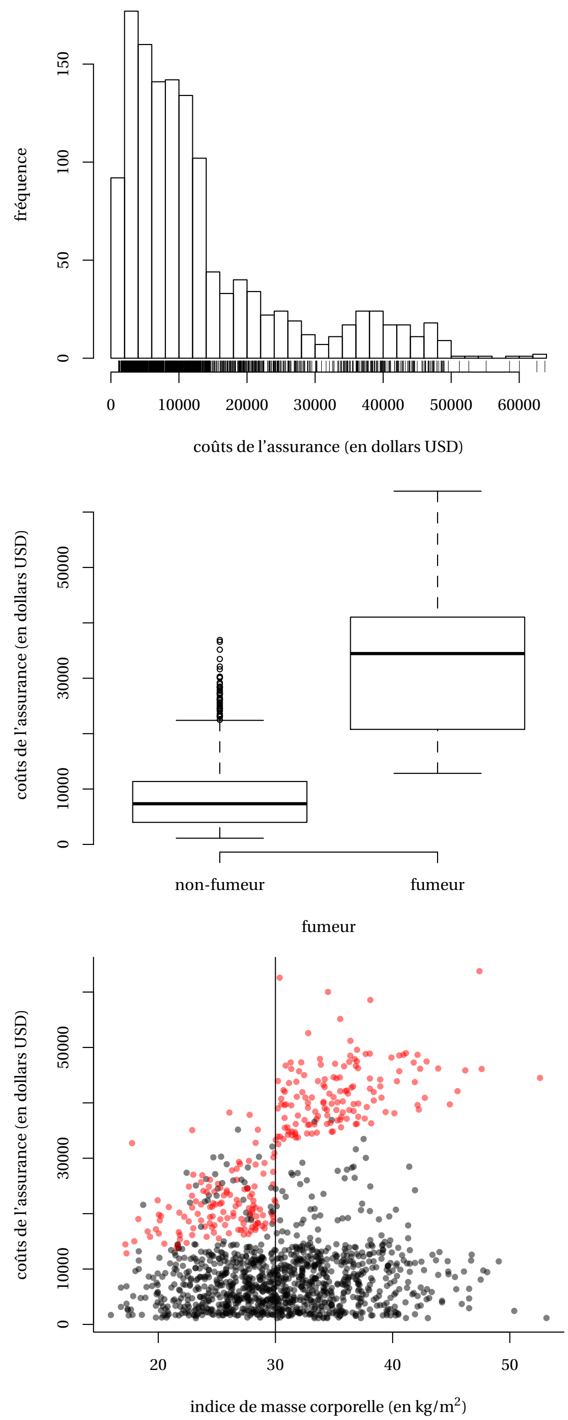

Figure 3: Histogramme des frais médicaux annuels (en dollars américains) (haut, gauche), boîte à moustache des frais en fonction du sexe (haut, gauche) et par statut de fumeur/non-fumeur (milieu, gauche). Nuages de points des frais contre l'âge (milieur, droit), qui montre une traîne linéaire pour trois 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).

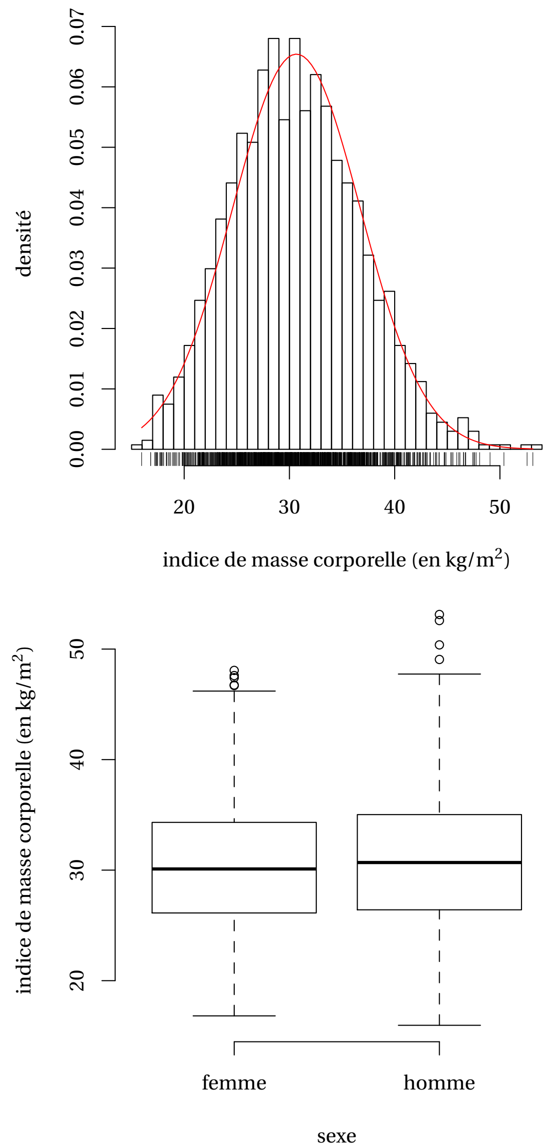

Figure 4: Histogramme de l'indice de masse corporelle (IMC), en kg/m^2, avec traits (haut, gauche), et boîte à moustache montrant l'interaction entre l'IMC et (a) le status de fumeur (haut, droit), (b) le sexe (bas, gauche) et (c) la région géographique de provenance (bas, droite).