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 , il faut utiliser l'option `sides=l` (c'est-à-dire "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.
(b) Testez cette hypothèse à l'aide de SAS et concluez.

2.2 **Influence des valeurs extrêmes dans le test-t** Le test-t est sensible à la présence de valeurs extrêmes dans les données, et c'est ce qui a motivé le test de la somme des rangs de Wilcoxon, 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.
(b) Identifiez la présence de la valeur extrême sur les boxplots et histogrammes des données fournis par la procédure `ttest`.
(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. Commentez sur la robustesse de ce test face aux valeurs extrêmes, en comparaison au test-t.

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)

Compiled 21/01/2020 at 15:19

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. Qu'observez-vous ?
(b) Écrivez formellement l'hypothèse statistique que vous voulez tester.
(c) Effectuez le test approprié, conclure et interpréter les résultats.
(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.
(e) Vérifiez graphiquement la normalité des données dans chaque groupe et concluez sur la validité de votre test.
(f) Reprendre les mêmes analyses que précédemment en utilisant cette fois un test de Wilcoxon. Comparez vos résultats avec ceux du test-t classique

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 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.

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.

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?
(b) Les fumeurs paient-ils des frais médicaux en moyenne équivalents aux non-fumeurs? Justifiez adéquatement votre réponse
(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.
(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 $\alpha = 0.05$.
(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?

Compiled 21/01/2020 at 15:19