Exercice 2 - Expérimentation statistique

Karim Derouiche

5/20/2022

Table des matières

Première partie

Création des données

Nous allons dans un premier temps créer le tableau de données:

```
actifs <- c(1150, 1500, 1700, 1800, 1800, 1850, 2200, 2700, 2900, 3000, 3100,
            3500, 3900, 4000, 5400)

etudiants <- c(300, 700, 850, 900, 1000, 1420, 1500, 1500, 1800, 1800, 1900,
               2200, 2400)

status <- c(rep("actif", 15), rep("etudiant", 13))

df = data.frame(status, recette = c(actifs, etudiants))
```

Question 1 : Intervalle de confiance de l'espérance μ

Les recettes suivent, comme supposé, une loi normale: $X \sim N\left(\mu, \sigma^2\right)$ et $X_1, \ldots, X_n$ n variables i.i.d selon la loi de X. Dans contexte idéal, l'intervalle de confiance au seuil de 5% de la moyenne serait de la forme:

$$IC = \left[\overline{X}_n \pm q\frac{\sigma}{\sqrt{n}}\right]$$

avec:

- $\overline{X}_n$: la moyenne empirique de notre échantillon
- q : le fractile d'ordre 0,975 de la loi normale centrée réduite
- σ : La variance connue
- n le nombre d'individus

Toutefois, dans notre situation la variance est inconnue. Nous devons alors l'estimer. Un estimateur sans biais de la variance serait:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$$

Nous avons donc:

$$\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \sim t(n-1) \quad \text{(loi de Student à } n-1 \text{ degrés de libertés).}$$

L'intervalle de confiance serait donc:

$$IC = \left[\bar{X}_n \pm t\frac{S_n}{\sqrt{n}}\right]$$

Où:

- t est le fractile d'ordre $1 - \frac{\alpha}{2}$ de la loi de Student à n-1 degré de liberté.

```
# Calduls des intervalles de confiance
me = mean(df$recette[df$status == "etudiant"])
ma = mean(df$recette[df$status == "actif"])
vare = var(df$recette[df$status == "etudiant"])
vara = var(df$recette[df$status == "actif"])
qte = qt(0.975, 12)
qta = qt(0.975, 14)
```

```
ICe = c(me-qte*(sqrt(vare)/sqrt(13)), me+qte*(sqrt(vare)/sqrt(13)))
ICa = c(ma-qta*(sqrt(vara)/sqrt(15)), ma+qta*(sqrt(vara)/sqrt(15)))
```

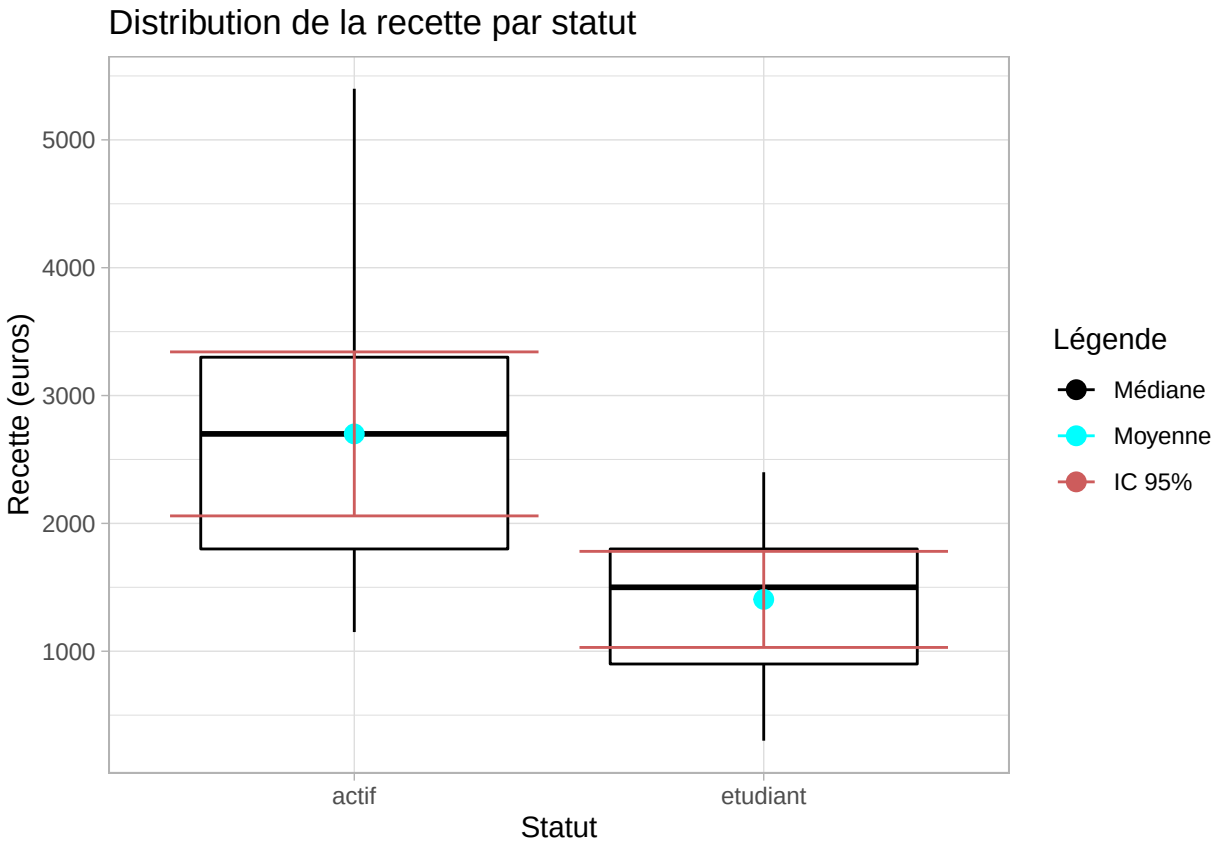
Pour les étudiants, l'intervalle de confiance de la moyenne est:

$$IC = [1030, 1781]$$

Pour les actifs, l'intervalle de confiance de la moyenne est:

$$IC = [2059, 3341]$$

Nous pouvons donc illustrer graphiquement nos résultats:



Question 2: Comparaison de moyennes à variances égales

Nous supposons que l'on a deux échantillons suivent une loi normale et où les variances sont connues.

Nous voulons tester, par le biais d'un test de Student:

$$H_0 : \mu_1 = \mu_2$$

contre

$$H1 : \mu_1 \neq \mu_2$$

La statistique de test est donnée par la formule suivante:

$$T = \frac{(\bar{X} - \bar{Y})}{S\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathcal{T}_{n_1+n_2-2}$$

Où notre estimateur de la variance non biaisé est tel que:

$$S^2 = \frac{(n_1 - 1)\, S_1^2 + (n_2 - 1)\, S_2^2}{n_1 + n_2 - 2}$$

La zone de rejet associée à notre test correspond à:

$$|T| \geq t_{1-\frac{\alpha}{2}}$$

```
S = sqrt((12*vare+14*vara)/(13+15-2))
Tstat = (me-ma)/(S*sqrt(1/13+1/15))
```

Or nous avons:

- $|T| = 3.5992$
- $t_{1-\frac{\alpha}{2}} = 2.055529$

Nous rejetons donc l'hypothèse nulle d'égalité des moyennes au seuil de confiance de 5%. Nous concluons en répondant à notre manager que, en moyenne, la recette d'un billet vendu sur ce vol, semble dépendre du statut étudiant vs actif.

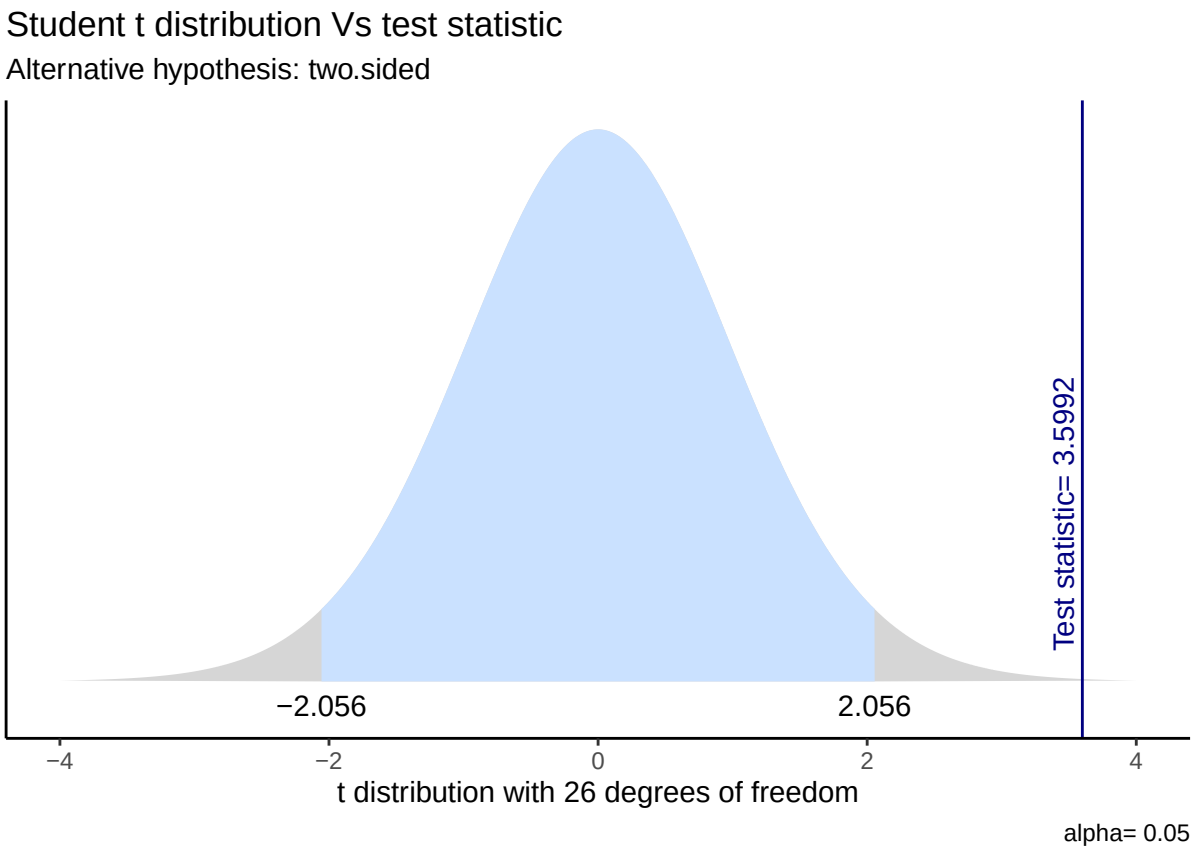
Question 3: Vérification du résulat

Pour vérifier notre résultat, on applique la fonction t.test sur notre jeu de données.

```
t.test(df$recette[df$status == "actif"], df$recette[df$status == "etudiant"], var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  df$recette[df$status == "actif"] and df$recette[df$status == "etudiant"]
## t = 3.5992, df = 26, p-value = 0.001317
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   555.2452 2033.9856
## sample estimates:
## mean of x mean of y
##  2700.000  1405.385
```

Visualisation de nos résultats:



Nous confirmons bien le rejet de l'hypothèse nulle au vue de la p-value.

Question 4: Comparaison de moyennes à variances non égales

Nous supposons que l'on a deux échantillons suivent une loi normale et où les variances sont inégales

Nous voulons toujours tester, cette fois, par le biais d'un test de Welch:

$$H_0 : \mu_1 = \mu_2$$

contre

$$H1 : \mu_1 \neq \mu_2$$

La statistique de test est donnée par la formule suivante:

$$T = \frac{(\bar{X} - \bar{Y})}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}}$$

Nous devons alors estimer le nombre de degré de liberté:

$$df = \left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2 / \left(\frac{S_1^4}{n_1^2\,(n_1 - 1)} + \frac{S_2^4}{n_2^2\,(n_2 - 1)}\right)$$

```
T_Welch = (me-ma)/(sqrt(vare/13+vara/15))
dfeedom = ((vare/13+vara/15)^2)/((vare^2/(13^2*12))+(vara^2/(15^2*14)))
```

La zone de rejet associée à notre test correspond à:

$$|T| \geq t_{1-\frac{\alpha}{2}}$$

Or nous avons:

- $|T| = 3.7597$
- $df = \sim 22$
- $t_{1-\frac{\alpha}{2}} = 2.073757$

Nous rejetons donc l'hypothèse nulle d'égalité des moyennes au seuil de confiance de 5%.

Nous pouvons vérifier nos résultats grâce à la fonction R:

```
t.test(df$recette[df$status == "actif"], df$recette[df$status == "etudiant"])
```

```
##
##  Welch Two Sample t-test
##
## data:  df$recette[df$status == "actif"] and df$recette[df$status == "etudiant"]
## t = 3.7496, df = 22.021, p-value = 0.001107
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   578.607 2010.624
## sample estimates:
## mean of x mean of y
##  2700.000  1405.385
```

Nous voyons bien que nous pouvons rejeter l'hypothèse nulle au seuil de 1%.

Question 5: Comparaison et conclusion des tests de moyennes

La recette générée par un billet Paris-Bangkok semble dépendre de manière significative du profil client. En effet, nous avons fait 2 tests d'hypothèses d'égalité de moyennes, les deux ont été rejetés. En conséquence il semble exister une réelle différence suivant à ce que le client soit un étudiant ou un actif.

Question 6: Comparaison de deux variances

Estimation des variances:

Comme nous l'avons lors de la question 1, un estimateur sans biais pour nos variance serait:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$$

Nous savons par ailleurs que, sous l'hypothèse de la normalité de la recette, la variance va suivre une loi du Khi deux:

$$S_n^2 \sim \frac{\sigma^2}{n-1}\chi^2_{(n-1)}$$

La loi du Khi deux n'était pas symétrique, nous obtenons alors la probabilité suivante:

$$\mathbb{P}\left(v_{\alpha/2} \leq \frac{n-1}{\sigma^2}S_n^2 \leq v_{1-\alpha/2}\right) = 1-\alpha$$

Ceci équivaut à

$$\mathbb{P}\left(\frac{(n-1)S_n^2}{v_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{v_{\alpha/2}}\right) = 1-\alpha.$$

Nous obtenons donc un intervalle de confiance pour la variance σ^2 avec seuil de confiance $1-\alpha$:

$$\left[\frac{(n-1)S_n^2}{v_{1-\alpha/2}}, \frac{(n-1)S_n^2}{v_{\alpha/2}}\right].$$

Pour les étudiants, l'intervalle de confiance de l'écart type est:

$$IC = [446, 1027]$$

Pour les actifs, l'intervalle de confiance de l'écart type est:

$$IC = [848, 1827]$$

Question 7: Egalité des variances

Test statistique

Nous voulons tester, avec un test de Fisher:

$$H_0 : \sigma_1 = \sigma_2$$

$$H1 : \sigma_1 \neq \sigma_2$$

La statistique de test est:

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$$

Le test s'effectue comme suit:

```r
var.test(df$recette[df$status == "actif"], df$recette[df$status == "etudiant"])
```

```
##
##  F test to compare two variances
##
## data:  df$recette[df$status == "actif"] and df$recette[df$status == "etudiant"]
## F = 3.4682, num df = 14, denom df = 12, p-value = 0.03713
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   1.081728 10.578698
## sample estimates:
## ratio of variances
##            3.46825
```

Nous rejettons l'hypothèse nulle d'égalité des variances au seuil de 5%.

Déductions:

Parmi les deux tests construits questions 2 et 4, le second est plus adéquat. En effet, nous ne pouvons pas supposer que les variances sont égales. Les deux sous populations ne semblent pas homogénènes étant donné l'hétéroscédasticité de la variance.

Intuitivement, nous pouvons supposer que les etudiants forment une classe de revenu plutôt faible, tandis que les actifs forment différentes classes de revenu. Ce qui leur permet parfois de s'éloigner du prix moyen pour un vol.

Deuxième partie

Question 1: Proposition de modèle

L'objectif est probablement de proposer les meilleurs vins aux clients. Toutefois, nous devons prendre compte certains facteurs, comme le client ou le contexte pour s'affranchir de certains biais lié à l'évaluation.

Nous pensons que le modèle se doit d'expliquer plus que de prédire l'évaluation de certains vins.

Un modèle linéaire peut très bien avoir un bon pouvoir explicatif. Nous pouvons par exemple imaginer un modèle de la sorte:

$$Note = \alpha + \sum_{i=1}^{i=6} \beta_i * \mathbf{1}_{Vin_i} + \sum_{i=1}^{i=3} \gamma_i * \mathbf{1}_{Context_i} + \sum_{i=1}^{i=12} \theta_i * \mathbf{1}_{Client_i}$$

Ce type de modèle pourrait nous permettre d'isoler un à un les effets de chaque facteurs.

Nous pouvons imaginer un second modèle, plus complet, qui mettrait en concurrence ce premier, où les interractions seraient également prises en compte:

$$Note = \alpha + \sum_{i=1}^{i=6} \beta_i * \mathbf{1}_{Vin_i} + \sum_{i=1}^{i=3} \gamma_i * \mathbf{1}_{Context_i} + \sum_{i=1}^{i=12} \theta_i * \mathbf{1}_{Client_i} + \sum_{i=1}^{i=18} \zeta_i * \mathbf{1}_{Vin:Context_i}$$
$$+ \sum_{i=1}^{i=72} \eta_i * \mathbf{1}_{Vin:Client_i} + \sum_{i=1}^{i=36} \rho_i * \mathbf{1}_{Contexte:Client_i} + \sum_{i=1}^{i=216} \phi_i * \mathbf{1}_{Vin:Contexte:Client_i}$$

Ce modèle isolerait également les effets croisés. Cependant il faudrait éventuellement être vigilant quant à la multicolinéarité et le nombre élevé de variables. C'est pourquoi la comparaison avec le modèle initial pourrait nous permettre de déceler les coefficients donnant la même information.

Question 2: L'interraction entre les facteurs de type vin et client

Concrètement, l'interraction entre les facteurs de type vin et client est caractérisée par les coefficients: $\sum_{i=1}^{i=72} \eta_i * \mathbf{1}_{Vin:Client_i}$.

Cela signifie que nous créons un coefficient pour chaque pair vin vs client. Comme nous avons 6 types de vin différents et 12 clients, nous avons donc 72 interractions différentes.

Question 3: L'interaction type de vin.client

Le test de significativité de l'interraction correspond à un test de significativité de Fisher Global:

Nous voulons tester:

$$H_0 : \eta_1 = ... = \eta_{72} = 0$$

$$H1 : \text{Au moins un des } \eta_i \neq 0$$

La statistique de test est telle que (toutes choses égales par ailleurs):

$$F = \frac{\text{SCE}/K}{\text{SCR}/(n-K-1)} = \frac{R^2/K}{\left(1-R^2\right)/(n-K-1)}$$

Ainsi, cette dernière suit une loi de Fisher:

$$F \sim Fisher(K, n-k-1)$$

Nous avons donc $pvalue = 0.005743$, nous pouvons donc rejeter l'hypothèse nulle de non significativité des coefficients de l'interraction vin.client au seuil de 1%.

Question 4: Effets à retenir

Concernant les effets, nous pouvons retenir le vin, le client et l'interraction vin.client. Ceux-ci sont significatifs au moins au seuil de 1%. Contrairement à la variable contexte, qui ne semble pas significatif.

Question 5: Vin le plus apprécié

Compte tenu des résulats et de la significativité, le vin 1 semble être le plus apprécié. Il s'agit du coefficient significatif le plus élévé (1.53). Toutefois, nous ne proposerions pas uniquement ce vin. En effet, il vaut mieux proposer une palette de choix car 12 clients ne peuvent pas être représentatifs de toute la population. Par ailleurs, la satifaction globale ne peut qu'augmenter si le client bénéficie d'un choix.

Dans le cas où la consommation est payante, nous pouvons éventuellement pondérer le prix à la qualité du vin.

Question 6: Interprétation des coefficients des clients

Le client 2 dispose d'un coefficient estimé à 0.06 mais ne semble pas être réellement significatif sur l'appréciation ($pvalue = 0.77 >> 0.05$). En d'autres termes, il n'est pas impactant dans notre modèle linéaire.

En revanche, le client 5 a un avis plus significatif ($pvalue < 0.01$). Celui-ci est par ailleurs plus sévère. En effet, indépendamment du contexte ou du vin, la note est impactée négativement de 0.72 points.