TP3: tests d'hypothèse et régression linéaire simple
Statistiques pour Sciences Humaines II

Baptiste Perez et Anna Kiriliouk

Année 2021–2022

NB : toutes les réponses ont été arrondis à deux chiffres après la virgule.

Réponses

1. 40.79% des individus de l'échantillon sont des femmes et 59.21% des individus de l'échantillon sont des hommes. Les hommes de l'échantillon ont généralement un QI un peu plus élevé que les femmes ; ceci se voit en calculant les moyennes (111.18 pour les hommes et 105.84 pour les femmes) ou en étudiant les boîtes à moustaches, qui montrent que les trois quartiles du QI sont (très) légèrement plus élevés pour les hommes que pour les femmes.
2. On obtient $(10.53, 13.15)$ pour les femmes et $(10.42, 12.83)$ pour les hommes. Ces intervalles sont très similaires ; celui des hommes est légèrement plus bas, c'est-à-dire, on s'attend à une note moyenne de la population des hommes légèrement plus basse que la note moyenne de la population des femmes.
3. Les hypothèses du test sont

$$H_0 : \mu_F = 110, \qquad H_1 : \mu_F \neq 110$$

   On trouve, sous l'hypothèse nulle,

$$\begin{aligned}\mu_F - \frac{s_F}{\sqrt{n_F}} t_{0.975} &= 110 - \frac{14.27}{\sqrt{31}} 2.04 = 104.77 \\ \mu_F + \frac{s_F}{\sqrt{n_F}} t_{0.975} &= 110 + \frac{14.27}{\sqrt{31}} 2.04 = 115.23\end{aligned}$$

   La région de rejet est

$$\{\bar{x} \leq 104.77 \text{ ou } \bar{x} \geq 115.23\}$$

   On a trouvé un QI moyen de 105.84. Cette valeur ne se trouve pas dans la région de rejet ; on ne peut pas rejeter l'hypothèse nulle (ou accepter l'hypothèse alternative). Il n'y a pas assez de preuves pour conclure que le QI moyen de la population des femmes n'est pas égal à 110 avec un niveau de confiance de 95%.
4. La probabilité de commettre une erreur de première espèce est égale au niveau de significativité du test ; ici,

$$\mathbb{P}[RH_0 \mid H_0] = \mathbb{P}[\bar{X} \leq 104.77 \text{ ou } \bar{X} \geq 115.23 \mid \mu_F = 110] = \alpha = 0.05$$

La probabilité de commettre une erreur de deuxième espèce se calcule par

$$
\begin{aligned}
\gamma &= \mathbb{P}[\cancel{RH}_0 \mid \mu_F = 105] = \mathbb{P}[104.77 \leq \bar{X} \leq 115.23 \mid \mu_F = 105] \\
&= \mathbb{P}\left[\frac{104.77 - 105}{14.27/\sqrt{31}} \leq T \leq \frac{115.23 - 105}{14.27/\sqrt{31}}\right] \\
&= \mathbb{P}[T \leq 3.99] - \mathbb{P}[T \leq -0.09] = 0.54
\end{aligned}
$$

Cette probabilité diminue au fur et à mesure que μ_F s'éloigne de la valeur de 110.

5. La p-valeur est la probabilité d'obtenir la même valeur de la statistique du test ou une valeur plus extrême (c'est-à-dire plus éloignée de l'hypothèse nulle), sous l'hypothèse nulle. On trouve

$$
\begin{aligned}
\mathbb{P}[\bar{X} \leq 105.84 \text{ ou } \bar{X} \geq 114.16 \mid H_0] &= 2\,\mathbb{P}[\bar{X} \leq 105.84 \mid H_0] \\
&= 2\,\mathbb{P}\left[T \leq \frac{105.84 - 110}{14.27/\sqrt{31}}\right] \\
&= 2\,\mathbb{P}\left[T \leq -1.62\right] = 0.1150
\end{aligned}
$$

On peut rejeter l'hypothèse nulle $H_0 : \mu_F = 110$ avec un "niveau de confiance" de maximum 88.5%.

6. Pour un test d'hypothèse unilateral, on aurait posé

$$H_0 : \mu_F = 110, \qquad H_1 : \mu_F < 110$$

La p-valeur serait alors

$$\mathbb{P}[\bar{X} \leq 105.84 \mid H_0] = \mathbb{P}\left[T \leq -1.62\right] = 0.0575$$

On peut rejeter l'hypothèse nulle $H_0 : \mu_F = 110$ avec un "niveau de confiance" de maximum 94.25%.

7. On fait un nuage de points entre `iq` et `grade_average`. Attention, on met **toujours** la variable explicative sur l'abscisse et la variable à expliquer sur l'ordonnée. On voit une association assez forte qui semble linéaire.
8. L'association est positive ; on peut calculer la corrélation et on trouve $r = 0.71$, donc l'association linéaire entre les deux variables est relativement forte. En moyenne, un QI plus élevé est associé avec une note moyenne plus élevée, et un QI plus bas est associé avec une note moyenne plus basse.
9. Voir fichier `R`.
10. On trouve $\widehat{\beta}_0 = -4.75$ et $\widehat{\beta}_1 = 0.15$, c'est-à-dire,

$$\texttt{grade\_average} = -4.75 + 0.15\,\texttt{iq}$$

On ne peut pas interpréter le coefficient $\widehat{\beta}_0$, car il représente la note moyenne à la quelle on s'attend (ici, une note négative !) pour un étudiant ayant un QI égal à 0. Il s'agit d'un extrapolation des données (plutôt que d'une interpolation).

11. On peut calculer

$$\begin{aligned}
\widehat{\mu}(100) &= -4.75282 + 0.15105 \times 100 = 10.35218\\
\widehat{\mu}(105) &= -4.75282 + 0.15105 \times 105 = 11.10743\\
\widehat{\mu}(108) &= -4.75282 + 0.15105 \times 108 = 11.56058\\
\widehat{\mu}(113) &= -4.75282 + 0.15105 \times 113 = 12.31583
\end{aligned}$$

Que le QI augmente de 100 à 105 points ou de 108 à 113 points, la note moyenne augmente, en moyenne, de 0.76 points. C'est la conséquence d'un modèle **linéaire** : si le QI moyen augmente d'un point, la note moyenne augmente de $\widehat{\beta}_1 = 0.15105$ points, et si le QI moyen augmente de cinq points, la note moyenne augmente de $5 \times \widehat{\beta}_1 = 5 \times 0.15105 = 0.75525$ points.

12. On trouve

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{s_{xy}}{s_x^2} = \frac{26.83}{177.6267} = 0.15105,\\
\widehat{\beta}_0 &= \bar{y} - \widehat{\beta}_1 \bar{x} = 11.71132 - 0.15105 \times 109 = -4.75282
\end{aligned}$$

13. Voir fichier `R`. *NB : si vous voulez exporter un graphe, il suffit de cliquer "Export" dans* `Rstudio` *(juste au dessus du graphe), et puis choisir "Save as image" ou "Copy to clipboard".*

14. On s'attend à une note de $-4.75282 + 0.15105 \times 74 = 6.42$. Or, l'étudiant (individu numéro 61) ayant un QI de 74 a obtenu une moyenne de 8.06. Il s'agit donc d'une sous-estimation (comme on aurait pu voir sur le graphe !) de 1.64 points, c'est-à-dire, le résidu $e_{61} = y_{61} - \widehat{y}_{61} = 8.06 - 6.42 = 1.64$. Ici, nous avons pu calculer le résidu parce qu'il existe un étudiant dans l'échantillon avec un QI de 74 ; on n'aurait pas pu calculer la même chose pour un QI de 75.