

Université de Caen
Normandie

IUT Grand Ouest Normandie

Bachelor Universitaire de Technologie
Science des Données
Campus de Lisieux

Science des Données 2 - Ressource 3.06
Tests d'hypothèses

Thématique

Étude sur les personnes en situation d'obésité

Auteurs

GAMONDELE Maxime

Année universitaire 2023-2024

Table des matières

Introduction	3
1 Étude du régime	4
1.1 Analyse exploratoire	4
1.2 Test statistique	5
2 Étude du poids	6
2.1 Analyse d'un échantillon de la population	6
2.2 Test statistique	7
Conclusion	8
Table des figures	9

Introduction

À travers cette étude, nous analysons un régime à base de fibres et un échantillon d'une population en surpoids. Dans le premier cas, nous utilisons une approche paramétrique, tandis que dans le second, nous adoptons une approche non paramétrique. Pour l'approche paramétrique, la distribution de notre population suit une loi connue, alors que dans l'approche non paramétrique, nous ne connaissons pas la loi sous-jacente à la distribution.

1 Étude du régime

Dans cette première partie, nous cherchons à tester quelles hypothèses conserver :

H_0 : Le poids reste identique avec le régime.

H_1 : Le poids diminue avec le régime.

Nous nous plaçons ici dans le cadre d'un test unilatéral inférieur.

1.1 Analyse exploratoire

Premièrement, pour avoir une idée de la distribution de notre échantillon, nous représentons la densité lissée de la variable Poids (cf. Fig. 1). Nous observons une cloche avec une légère dissymétrie, et la distribution semble unimodale. Nous décidons d'évaluer la possibilité que notre population suive une loi de Laplace-Gauss.

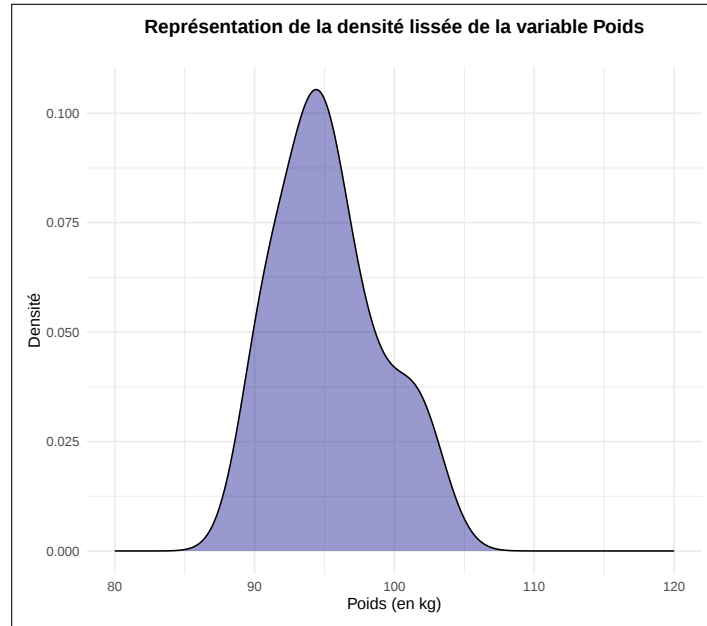


Fig. 1. Densité lissée de la variable Poids 1

Pour obtenir plus d'informations sur la distribution, nous représentons le Q-Q plot (cf. Fig. 2), ce qui permet de comparer les quantiles théoriques de la loi normale avec les quantiles observés de notre variable (ici, le poids). Les points semblent se rapprocher d'une droite, indiquant que notre échantillon suit probablement une loi gaussienne.

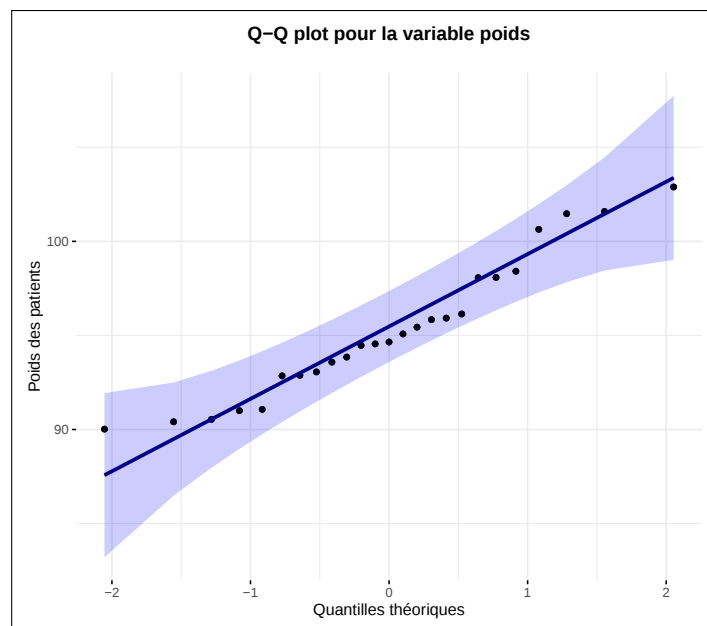


Fig. 2. Q-Q plot des valeurs centrées réduites 1

Après avoir visualisé ces graphiques, nous avons encore quelques doutes sur la distribution de notre échantillon. Nous effectuons donc un test de normalité de Shapiro-Wilk, dont les hypothèses sont :

$$H_0 : \text{La distribution de } X \text{ est gaussienne.}$$

$$H_1 : \text{La distribution de } X \text{ n'est pas gaussienne.}$$

Avec les résultats obtenus (cf. Fig. 3), nous pouvons affirmer que la distribution du poids semble suivre une loi de Laplace-Gauss avec $W = 0.95$ et une p-value de 0.192.

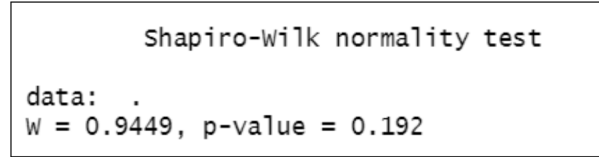


Fig. 3. Test de Shapiro-Wilk 1

1.2 Test statistique

Maintenant que l'hypothèse de normalité est validée, nous mettons en œuvre le test d'hypothèses dans un cadre paramétrique. Notre résultat graphique montre que nous pouvons rejeter l'hypothèse H_0 , car notre statistique de test T_{obs} est inférieure à la valeur critique $-t_{1-\alpha}$ (cf. Fig. 4). De plus, la p-value est presque égale à 0 (cf. Fig. 5).

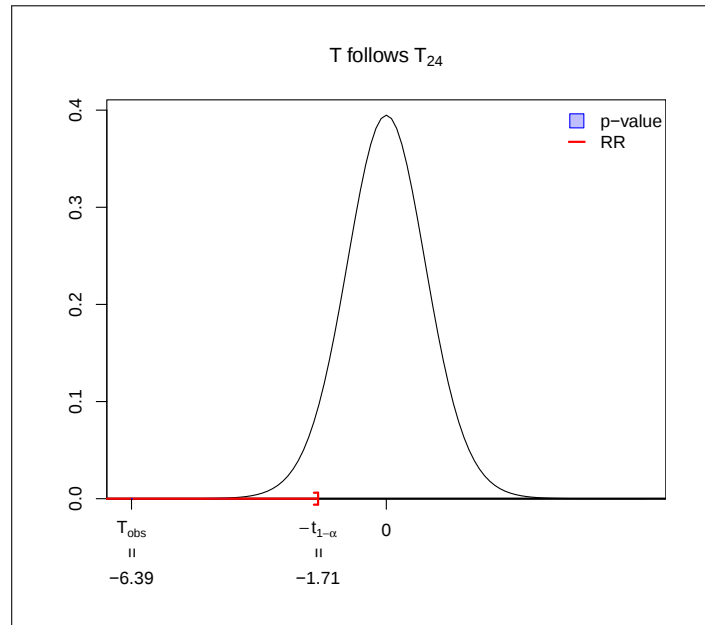


Fig. 4. Zone de rejet

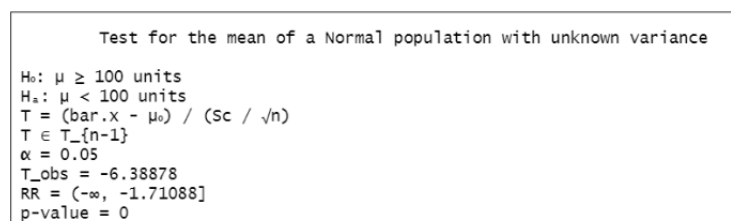


Fig. 5. Résultats du Mean.test

Pour conclure, nous pouvons affirmer que le régime à base de fibres pendant 6 mois permet une perte de poids.

2 Étude du poids

Nous étudions le poids d'une population à travers un échantillon de 40 individus. Nous confrontons nos hypothèses suivantes :

$$\begin{aligned}H_0 &: \text{Le poids moyen de notre population est égal à 100 kg.} \\H_1 &: \text{Le poids moyen de notre population est supérieur à 100 kg.}\end{aligned}$$

Nous nous plaçons ici dans le cadre d'un test unilatéral supérieur.

2.1 Analyse d'un échantillon de la population

Nous étudions la densité lissée de la variable Poids. La visualisation montre une courbe en cloche, mais cela ne suffit pas pour conclure que la distribution suit une loi de Laplace-Gauss (cf. Fig. 6).

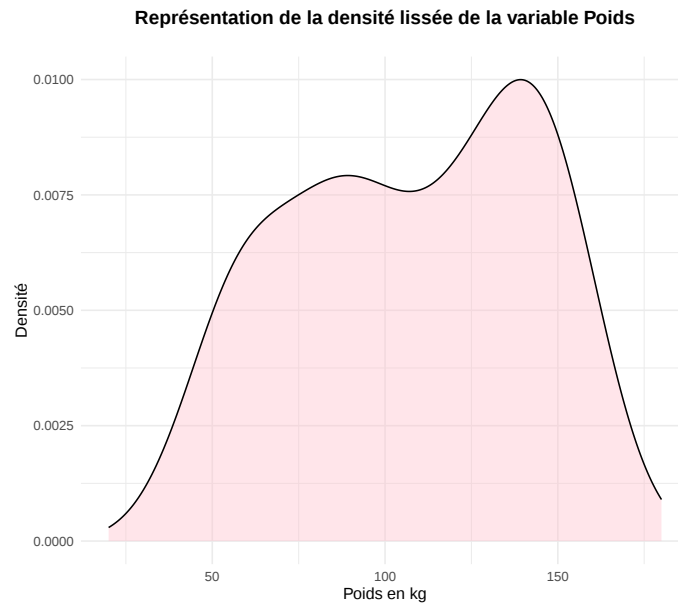


Fig. 6. Densité lissée de la variable Poids 2

De ce fait, nous décidons de visualiser le Q-Q plot (cf. Fig. 7). Nous remarquons que nos points semblent s'aligner sur une droite, ce qui signifie que notre échantillon se rapproche d'une loi normale.

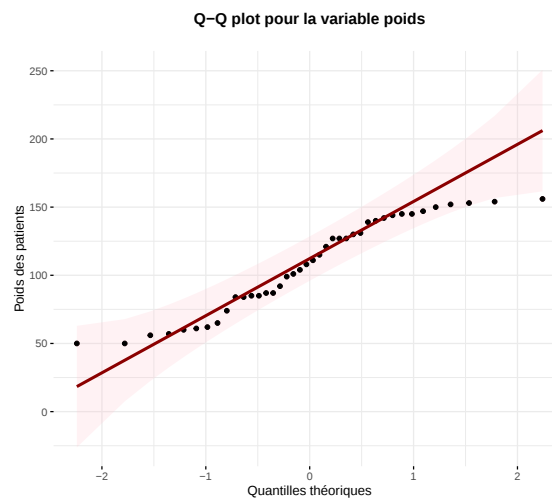


Fig. 7. Q-Q plot des valeurs centrées réduites 2

Enfin, nous effectuons le test de normalité de Shapiro-Wilk qui nous donne les résultats suivants (cf. Fig. 8). Ici, les résultats du test de Shapiro-Wilk suggèrent que notre échantillon n'est pas normalement distribué au niveau de signification de 5%. Par conséquent, nous rejetons l'hypothèse H_0 .

```

Shapiro-Wilk normality test

data: .
W = 0.92274, p-value = 0.009401

```

Fig. 8. Test de Shapiro-Wilk 2

2.2 Test statistique

Étant donné que notre échantillon ne suit pas une loi normale, nous allons utiliser l'approche non paramétrique, basée sur le théorème de la limite centrale (cf. Fig. 9). Notre statistique de test est $W = 1.408$. On note qu'elle est inférieure à la valeur critique, qui est de 1.644854. Étant donné que nous sommes dans un test unilatéral supérieur, la statistique de test n'atteint pas la zone de rejet de l'hypothèse nulle. La p-value, de 0.0795, est supérieure à α qui est égal à 0.05. Cela signifie qu'il n'y a pas suffisamment de preuves pour rejeter l'hypothèse nulle à un niveau de signification de 5

	W	Critical_value	p_value
1	1.408487	1.644854	0.07949349

Fig. 9. TCL

Nous proposons une alternative à ce test d'hypothèse. Au vu de la taille de notre échantillon ($n=40$), nous nous proposons d'approximer nos résultats à l'aide de la loi gaussienne, car notre échantillon est suffisamment grand $n > 30$. Nous proposons donc de réaliser un t-test dont les résultats sont présentés ci-dessous (cf. Fig. 10).

```

One Sample t-test

data: .
t = 1.4085, df = 39, p-value = 0.08345
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
 98.49393      Inf
sample estimates:
mean of x
 107.675

```

Fig. 10. t-test

De ce fait, nous pouvons conclure que la moyenne théorique est bien égale à 100 kg.

Conclusion

Dans le cadre de l'analyse paramétrique, nous notons le rejet de l'hypothèse H_0 .

Dans le cadre du second exercice, notre objectif était de déterminer les distributions de probabilité les mieux adaptées à chaque échantillon. En utilisant la même méthodologie, nous avons entrepris cette recherche. Les résultats de nos investigations ont abouti à l'identification des distributions suivantes : pour l'échantillon n°1, la loi exponentielle ; pour l'échantillon n°2, la loi uniforme ; pour l'échantillon n°3, la loi gaussienne ; et enfin, pour l'échantillon n°4, la loi lognormale.

À travers le troisième exercice, nous nous sommes demandés si la taille de l'échantillon avait une influence sur la qualité de l'estimateur. D'après nos simulations, nous constatons que pour un échantillon trop petit, le sommet de la courbe de densité est proche de la statistique de la population, mais la distribution est asymétrique. Nous avons aussi remarqué que lorsque la taille de l'échantillon était trop grande, ici 50, le sommet de la courbe de densité était éloigné de la vraie moyenne, mais avec une distribution presque symétrique. Nous pouvons donc proposer l'hypothèse qu'il faut que l'échantillon soit supérieur ou égal à 10, mais inférieur à 50 afin d'avoir une estimation la plus proche de la réalité.

Table des figures

1	Densité lissée de la variable Poids 1	4
2	Q-Q plot des valeurs centrées réduites 1	4
3	Test de Shapiro-Wilk 1	5
4	Zone de rejet	5
5	Résultats du Mean.test	5
6	Densité lissée de la variable Poids 2	6
7	Q-Q plot des valeurs centrées réduites 2	6
8	Test de Shapiro-Wilk 2	7
9	TCL	7
10	t-test	7