Faculté de Droit & des Sciences Économiques — Université de Limoges

ANNÉE UNIVERSITAIRE 2021-2022

Session 1 **Semestre 4**

Licence Economie-Gestion – 2e année

Matière : Statistiques et probabilités - Éléments de correction **Durée :** 2 heures
Enseignant : Vincent Jalby

Problème

En 2019, on évalue à 1602 € la dépense annuelle moyenne d'énergie par les ménages français pour leur logement (source : `http://www.statistiques.developpement-durable.gouv.fr`) avec un écart-type de 450 €. (Dans la suite, on notera simplement **dépense énergétique**.)

Partie I (30 min, 6 points)

Afin d'actualiser ces données pour l'année 2021, une collectivité territoriale a effectué un sondage sur un échantillon aléatoire de ménages. Les résultats sont disponibles dans la sortie JAMOVI (`https://www.jamovi.org`) ci-dessous.

Descriptives

				90% Confidence Interval				
	N	Mean	SE	Lower	Upper	SD	Minimum	Maximum
Depense	256	1655.793	29.051	1608.009	1703.577	464.811	185.000	2918.000

1) La population est l'ensemble des ménages français (ou, plus précisément, l'ensemble des ménages administrés par la collectivité). Le caractère étudié correspond à la dépense annuelle d'énergie du ménage et peut être considéré comme une variable aléatoire X de loi inconnue. Son espérance $\mathrm{E}(X) = \mu$ représente la dépense énergétique moyenne des ménages de la population et $\sigma = \sigma_X$ l'écart-type de cette dépense dans la population. Le sondage est modélisé par un n-échantillon $(X_1, \dots, X_n)$ correspondant à $n = 256$ variables aléatoires indépendantes et de même loi que X.

2) La statistique $\overline{X} = \frac{1}{n}\sum X_i$ est une estimateur sans biais et convergent de $\mu = \mathrm{E}(X)$. Une estimation de la dépense énergétique moyenne (de la population) est donc donnée par une observation de $\overline{X}$, c'est-à-dire, $\bar{x}$ correspondant à la moyenne (empirique) de l'échantillon. D'après la sortie JAMOVI ci-dessus, $\bar{x} = 1655.79$.

3) L'écart-type σ_X de la population étant inconnu, et l'échantillon étant de taille suffisante ($n = 256 > 100$), on utilise la statistique

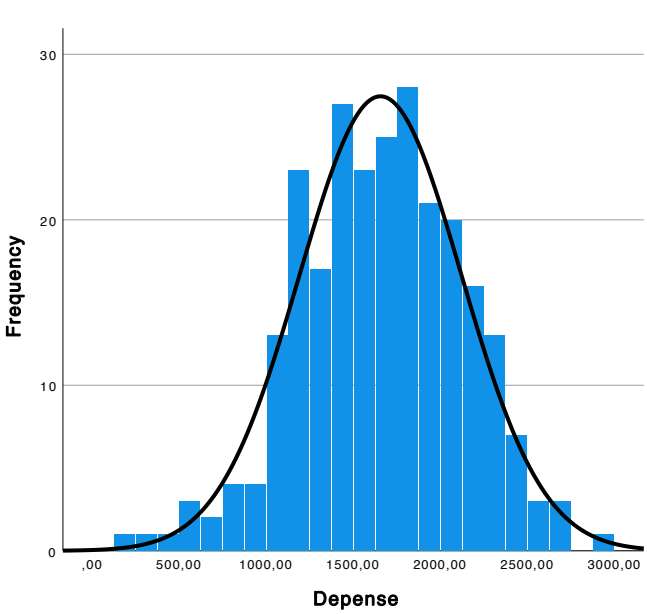
$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \underset{\sim}{\hookrightarrow} St(n-1)$$

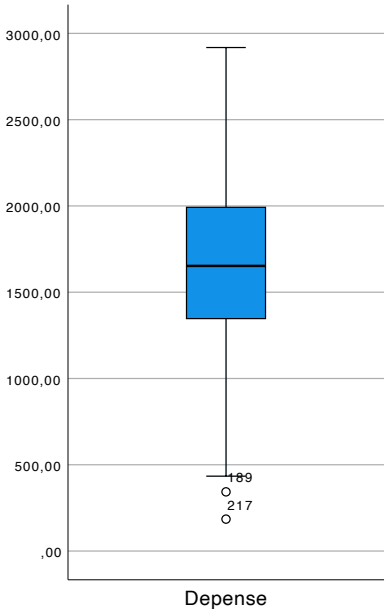
L'intervalle de confiance à 95 % est donc

$$\mathrm{IC}_{0.95}(\mu) = \left[\overline{X} \pm t_{0.975}\frac{S}{\sqrt{n}}\right] \implies \mathrm{ic}_{0.95}(\mu) = \left[1655.793 \pm 1.96\frac{464.811}{\sqrt{256}}\right] = [1598.85, 1712.73]$$

4) Sur un échantillon de $n = 256$ ménages, la dépense énergétique varie de 185 € à 2918 € avec une moyenne $\bar{x} = 1655.79$ € et un écart-type $s = 464.81$ €. La précision de l'estimation de la moyenne est donnée par l'erreur standard (se = 29.051). L'intervalle de confiance de la moyenne à 90 % ($\mathrm{ic}_{0.90}(\mu) = [1608, 1703.57]$) est plus précis (moins large) que celui trouvé à la question précédente, en raison d'un niveau de confiance moins élevé (90 % vs 95 %).

5) Aucune hypothèse n'ayant été faite sur la loi de X, il est nécessaire de vérifier quelle peut être supposée normale. Trois *vérifications* sont possibles à l'aide des sortie SPSS ci-dessous :





Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Depense		,030	256	,200*	,995	256	,488

Tout d'abord, on voit que l'histogramme de l'échantillon a approximativement la forme de la densité d'une loi normale. En outre, la boite à moustache est parfaitement symétrique. Finalement, le test de normalité de Shapiro-Wilk a une probabilité critique de 49 %, confirmant qu'il n'est pas possible de rejeter l'hypothèse de normalité de X. En conclusion, on peut considérer que l'échantillon étudié est bien issu d'une variable X suivant une loi normale.

Partie II (30 min, 5 points)

On souhaite d'abord déterminer si les variations de dépenses énergétiques ont changé en 2021 (par rapport à 2019).

1) Les variations des dépenses énergétiques sont modélisées par la variance (ou l'écart-type) de X. Il est donc nécessaire d'effectuer le test $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma \neq \sigma_0$ où $\sigma_0 = 450$ correspondant à l'écart-type de 2019. Un test bilatéral est choisi car on souhaite vérifier la différence (« ont changé »).

2) Comme on a vu que X suit une loi normale, on utilise la variable de décision

$$K^2 = \frac{(n-1)S^2}{\sigma_0^2} \hookrightarrow \chi^2(n-1) \implies k^2 = \frac{(255)464.811^2}{450^2} \approx 272.06$$

La région critique bilatérale est donc

$$W = [0, k^2_{0.05}] \cup [k^2_{0.95}, +\infty] \approx [0, 213.9] \cup [288.4, +\infty]$$

La table du χ^2 fournie s'arrêtant à $n = 200$, on estime la table pour $n = 255$ par celle de $n = 250$ en faisant une « interpolation linéaire » : $k^2_{0.05}(250) \approx k^2_{0.05}(200) + \left(k^2_{0.05}(200) - K^2_{0.05}(150)\right)$. Comme $k^2 = 272.06 \notin W$, on ne peut par rejeter l'hypothèse H_0 avec un risque de 10 %.

3) Les variations de dépenses énergétiques ne sont pas significativement différentes en 2021 de celles de 2019. Le risque pris est le risque de seconde espèce, $\beta = P(H_1|H_0)$ qui n'est pas calculé.

4) La sortie Stata correspond au test sur la variance précédemment effectué.

```
Stata. sdtest Depense == 450

One-sample test of variance
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 Depense |     256    1655.793    29.05067    464.8107    1598.583    1713.003
------------------------------------------------------------------------------
    sd = sd(Depense)                                          c = chi2 = 272.0617
Ho: sd = 450                                     degrees of freedom =      255

    Ha: sd < 450                 Ha: sd != 450                 Ha: sd > 450
  Pr(C < c) = 0.7789         2*Pr(C > c) = 0.4421          Pr(C > c) = 0.2211
```

On retrouve la valeur de la variable de décision (`c = chi2 = 272.0617`). La probabilité du test bilatéral (au centre) est donc 0.4421. Cette probabilité est largement supérieure au risque de première espèce de 10 %, on ne peut donc pas rejeter l'hypothèse nulle. Les variations de dépenses n'ont pas changé.

Partie III (30 min, 5 points)

On souhaite à présent tester si la dépense énergétique moyenne a augmenté de 2019 à 2021.

1) Il s'agit d'un test sur une moyenne. Le test est unilatéral à droite puisqu'on étudie l'augmentation des dépenses. Les hypothèses sont donc $H_0 : \mu = \mu_0$ contre $H_1 : \mu > \mu_0$ où $\mu_0 = 1602$ correspond à la dépense énergétique moyenne en 2019.
L'écart-type n'étant pas connu et l'échantillon étant de grande taille, on utilise la variable de décision

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \underset{\sim}{\hookrightarrow} St(n-1) \implies t = \frac{1655.792 - 1602}{464.811/\sqrt{256}} \approx 1.85$$

La région critique est $W =]t_{0.95}, +\infty[=]1.645, +\infty[$. Comme $t \in W$, on rejette l'hypothèse H_0 et on accepte l'hypothèse H_1.

2) La dépense énergétique moyenne a donc augmenté en 2021 par rapport à 2019. Le risque pris est le risque de première espèce $\alpha = P(H_1|H_0) = 5\ \%$.

3) La probabilité critique associée au test est

$$p\text{-}value = P(T > 1.85) = 1 - P(T < 1.85) \approx 1 - 0.9678 \approx 0.0322 \approx 3\ \ \%$$

(Pour ce calcul, comme n est très grand, on approche la table de $St(256)$ par celle de $\mathcal{N}(0,1)$ plus précise.)
Comme la *p-value* est inférieure au risque de première espèce (de 5 %), on retrouve le rejet de l'hypothèse nulle.

4) En prenant un risque de 1 %, on ne peut pas rejeter l'hypothèse nulle car $p\text{-}value = 3\ \% \not< 1\ \%$.

5) La sortie R correspond à un test T bilatéral.

```
R> t.test(depense, mu=1602)

        One Sample t-test

data:  depense
t = 1.8517, df = 255, p-value = 0.06523
alternative hypothesis: true mean is not equal to 1602
95 percent confidence interval:
 1598.583 1713.003
sample estimates:
mean of x
 1655.793
```

On retrouve la valeur de la variable de décision `t = 1.8517` ainsi que la probabilité critique bilatérale `p-value = 0.06523`.
On obtient la probabilité critique unilatérale en divisant la précédente par deux, soit $p\text{-}value_R \approx 3.3\ \%$.

Partie IV (30 min, 4 points)

L'échantillon précédent est composé de ménages résidant en habitat collectif (immeuble) et individuel (maison). On souhaite déterminer si la dépense énergétique moyenne est différente selon le type d'habitation.

1) Dans cette partie, on divise la population en deux (individuel et collectif). Cela revient à considérer X la dépense énergétique des (ménages résidant dans des) logements de type individuel et Y celle des logements collectifs. On note μ_X et σ_X (resp. μ_Y et σ_Y) la moyenne (sur la population) et l'écart-type des dépenses énergétiques pour les logements individuels (resp. collectifs).
Pour répondre à la question de cette partie, on doit tester $H_0 : \mu_X = \mu_Y$ contre $\mu_X \neq \mu_Y$ ou plus précisément, $H_0 : \mu_X - \mu_Y = 0$ contre $\mu_X - \mu_Y \neq 0$.
Avant cela, on procèdera au test d'égalité des variances $H_0 : \sigma_X = \sigma_Y$ contre $\sigma_X \neq \sigma_Y$.

2) Les échantillons étudiés sont composés de 199 logements individuels et 137 logements collectifs.

Type	Method	N	Mean	Std Dev	Std Err	Mean	95% CL Mean		Std Dev
Individuel		119	1658.8	438.5	40.1990	1658.8	1579.2	1738.4	438.5
Collectif		137	1562.3	450.4	38.4807	1562.3	1486.2	1638.4	450.4
Diff (1-2)	Pooled		96.4889	444.9	55.7533	96.4889	-13.3088	206.3	444.9
Diff (1-2)	Satterthwaite		96.4889		55.6482	96.4889	-13.1087	206.1	

On observe (sur les échantillons) une dépense énergétique moyenne supérieure pour les logements individuels ($\bar{x} = 1658.8$ vs $\bar{y} = 1562.3$). Toutefois, les variations de dépenses sont supérieures pour les logements collectifs ($s_Y = 450.4$ vs $s_X = 438.5$).

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	136	118	1.05	0.7675

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	254	1.73	0.0847
Satterthwaite	Unequal	250.7	1.73	0.0842

Le test d'égalité des variances à une probabilité critique (bilatérale) de plus de 76 %, bien supérieure aux risques de premier espèce habituels (5 ou 10 On en conclut qu'on ne peut pas rejeter l'hypothèse nulle d'égalité des variances. Donc, dans la suite, on supposera que $\sigma_X = \sigma_Y$.

Le test d'égalité des moyennes (ligne *Pooled* puisque les variances sont supposées égales) a une probabilité critique égale à 8.47 %.

Si on prend un risque habituel de $\alpha = 5$ %, alors on ne peut pas rejeter l'hypothèse nulle : les données collectées ne permettent pas de conclure que la dépense énergétique moyenne est **différente** selon le type d'habitation avec un risque de 5 %.

Si, on avait choisi d'effectuer un test unilatéral, en prenant comme hypothèse alternative l'hypothèse (naturelle) $\mu_X > \mu_Y$, alors la probabilité critique du test serait $8.47/2 = 4.2\ \% < 5\ \%$. On rejetterait donc H_0 et on conclurait que la dépense énergétique moyenne des logements individuels est **supérieure** à celles des logements collectifs.