Faculté de Droit & des Sciences Économiques — Université de Limoges

ANNÉE UNIVERSITAIRE 2024-2025

Session 1 **Semestre 4**

Licence Economie-Gestion – 2e année

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

Calculatrices non-programmables et non graphiques autorisées. Aucun document autorisé.

Problème

Depuis le 1er janvier 2019, la distribution de l'eau potable aux usagers de 9 communes (dont Limoges) de l'agglomération de Limoges est assurée en régie par la communauté urbaine de Limoges-Métropole.

En 2023, le volume d'eau facturé en vente directe à l'abonné s'élevait à 8 103 329 m^3 pour 166 395 habitants desservis, soit environ 133 litres par jour et par habitant, avec un écart-type de 40 litres.

Source : *Limoges-Métropole, Rapport annuel 2023 sur le prix et la qualité du service public de l'eau, publié le 15 juillet 2024.* `www.limoges-metropole.fr`

En attendant la publication des chiffres officiels de 2024, une association souhaite estimer la consommation moyenne pour 2024. Pour cela, elle a relevé la consommation annuelle pour 2024 de 151 foyers de l'agglomération choisis au hasard.

Partie I (15 min, 3 points)

Les données (consommation en litres par jour et par habitant) collectées sur l'échantillon sont résumées dans la sortie jamovi suivante :

Descriptives

									Shapiro-Wilk	
	N	Missing	Mean	SE	Median	SD	Minimum	Maximum	W	p
Conso	151	0	139.0	3.406	139.0	41.86	26.00	256.0	0.9963	0.973

1) On étudie, sur la population (Ω) des habitants (ou des ménages) de Limoges-Métropole, la consommation d'eau potable (par jour et par habitant) représentée par la variable X.
On note $\mu = \mathrm{E}(X)$ la consommation moyenne et $\sigma = \sigma_X$ son écart-type, les deux étant inconnus (en 2024).
Pour cela, on utilise un 151-échantillon $(X_1, \dots, X_{151})$ correspondant à 151 variables aléatoires indépendantes et de même loi que X.

2) Sur les $n = 151$ consommations relevées qui vont de 26 à 256 litre, la consommation moyenne (par jour et par habitant) est de $\bar{x} = 139$ litres avec un écart-type de $s = 41.86$ litres. L'erreur standard sur la moyenne (donnant la précision de l'estimation de μ par $\bar{x}$) est de 3.4 litres. Finalement, le test de normalité de Shapiro-Wilk, avec une *p-value* $= 0.973$, ne permet pas de rejeter l'hypothèse de normalité de X. On pourra donc supposer dans la suite que $X \sim \mathcal{N}(\mu, \sigma^2)$.

3) L'écart-type σ étant inconnu (en 2024), on utilise la statistique

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim St(n-1) \implies \mathrm{IC}_{0.90}(\mu) = \left[\overline{X} \pm t_{0.95}\frac{S}{\sqrt{n}}\right] \implies \mathrm{ic}_{0.90}(\mu) = \left[139 \pm 1.655\frac{41.86}{\sqrt{151}}\right] = [133.36, 144.64]$$

où $t_{0.95} = 1.655$ est le quantile d'ordre 95 % de la loi $St(150)$.

Partie II (25 min, 4 points)

On souhaite vérifier que les variations de consommation n'ont pas augmenté par rapport à 2023.

1) Il s'agit d'un test sur une variance. L'hypothèse nulle est l'hypothèse de stabilité soit $H_0 : \sigma = \sigma_{2023} = 40$. L'hypothèse alternative correspond à une augmentation des variations, soit $H_1 : \sigma > 40$.

2) Comme on suppose que X suit une loi normale (test de Shapiro-Wilk, Partie I), on a

$$K^2 = \frac{(n-1)S^2}{\sigma_{2023}^2} \sim \chi^2(n-1) \implies k^2 = \frac{150 \times 41.86^2}{40^2} = 164.27$$

S'agissant d'un test unilatéral à droite ($H_1 : \sigma > 40$), la région critique est $W =]k^2_{0.95}, +\infty[=]179.6, +\infty[$. Comme $k^2 = 164.27 \notin W$, on ne peut pas rejeter l'hypothèse H_0.

3) Les données collectées ne permettent donc pas de conclure que les variations des consommations ont augmenté. Le risque pris est le risque de seconde espèce $\beta = P(H_0|H_1)$ qui n'est pas évalué.

4) La sortie R correspond à un test bilatéral sur la variance. On retrouve la valeur de la variable de décision (`X-squared` $= 164.25$) mais aussi la probabilité critique bilatérale associée au test ($p\text{-}value_B = 0.3623$). En divisant par deux, on trouve la probabilité critique unilatérale (correspondant au test de l'exercice) $p\text{-}value_U = 0.1812$. Celle-ci étant supérieure au risque fixé (5 %), elle confirme qu'il n'est pas possible de rejeter l'hypothèse nulle. En outre, il faudrait prendre un risque de près de 20 % pour pouvoir conclure à une augmentation des variations !

```
R> VarTest(data$Conso, sigma.squared = 40**2)

One Sample Chi-Square test on variance

data:  data$Conso
X-squared = 164.25, df = 150, p-value = 0.3623
alternative hypothesis: true variance is not equal to 1600
95 percent confidence interval:
 1414.431 2227.427
sample estimates:
variance of x
     1752.013
```

Partie III (30 min, 5 points)

On souhaite à présent tester si la consommation moyenne a augmenté en 2024 (par rapport à 2023).

1) Il s'agit d'un test sur une moyenne. En 2023, la consommation moyenne était $\mu_0 = 133$ litres. L'hypothèse alternative, correspondant à une hausse, est donc $H_1 : \mu > 133$. L'hypothèse nulle, correspondant à une stabilité de la consommation, est $H_0 : \mu = 133$. Il s'agit donc d'un test unilatéral à droite.
La variable de décision est

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim \text{St}(n-1) = \text{St}(150) \implies t = \frac{139 - 133}{3.406} = 1.76$$

Avec un risque de première espèce de 5 %, la région critique (à droite) est $W =]t_{0.95}, +\infty[=]1.655, +\infty[$. Comme $t \in W$, on rejette l'hypothèse H_0 et on accepte l'hypothèse H_1.

2) Le test conclut donc à une augmentation significative de la consommation moyenne d'eau en 2024 par rapport à 2023. Le risque pris est le risque de première espèce $\alpha = P(H_1|H_0) = 5$ %.

3) La probabilité critique associée à ce test est $p\text{-}value = P(T > 1.76) \approx 1 - 0.96 = 4$ %. Comme $p\text{-}value < 5$ %, cela confirme qu'on rejette l'hypothèse H_0 pour accepter l'hypothèse H_1.

4) La sortie STATA correspond au test précédent. Outre les statistiques descriptives de l'échantillon, on retrouve la valeur de la variable de décision ($t = 1.7692$) ainsi que les probabilités critiques des tests. Celle concernant notre test (unilatéral à droite) est `Pr(T > t) = 0.0394` proche des 4 % trouvés ci-dessus.

```
Stata. ttest Conso == 133

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
   Conso |     151    139.0265    3.406278    41.85705     132.296     145.757
------------------------------------------------------------------------------
    mean = mean(Conso)                                            t =   1.7692
H0: mean = 133                                 Degrees of freedom =      150

   Ha: mean < 133               Ha: mean != 133                 Ha: mean > 133
 Pr(T < t) = 0.9606         Pr(|T| > |t|) = 0.0789          Pr(T > t) = 0.0394
```

Partie IV (25 min, 4 points)

Les ménages sondés dans l'enquête précédente étaient répartis sur l'ensemble du territoire de Limoges-Métropole (Limoges et communes limitrophes). On souhaite comparer la consommation moyenne des habitants de Limoges à ceux des autres communes.

La sortie SAS suivante précise les consommations d'eau des habitants de Limoges et celles des habitants des autres communes de l'échantillon.

Variable: Conso

Commune	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Limoges		120	142.9	41.2925	3.7695	58.0000	256.0
Autres communes		31	124.2	41.3737	7.4309	26.0000	191.0
Diff (1-2)	Pooled		18.6242	41.3088	8.3226		
Diff (1-2)	Satterthwaite		18.6242		8.3323		

1) La population de LM est à présent décomposée en deux sous-population Ω_1 et Ω_2. On note X la consommation des habitants de Limoges (sur Ω_1) et Y celle des habitants des autres communes de LM (sur Ω_2). On suppose que les deux variables sont normales et on note μ_X, σ_X (resp. μ_Y, σ_Y) la moyenne et l'écart-type de X (resp. de Y) sur chacune des sous-populations. Les données collectées correspondent donc à deux échantillons $(X_1, \dots, X_{120})$ pour celui de Limoges et $(Y_1, \dots, Y_{31})$ pour celui des autres communes.

La sortie SAS montre que la consommation moyenne de l'échantillon de Limoges ($\bar{x} = 142.9$) est près de 20 litres supérieure à celle de l'échantillon des autres communes ($\bar{y} = 124.2$). Toutefois, l'erreur-standard de la moyenne (précision de l'estimation) est double (7.42 vs 3.77) sur le second échantillon, en particulier en raison de sa faible taille (31).

2) Pour affirmer que la consommation moyenne des habitants de Limoges est supérieure à celles des autres communes, nous devons effectuer un test T sur deux échantillons indépendants avec les hypothèses $H_0 : \mu_X = \mu_Y$ vs $H_1 : \mu_X > \mu_Y$. Avant cela, nous devons tester l'égalité des variances ($H_0 : \sigma_X = \sigma_Y$) à l'aide d'un test F.

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	30	119	1.00	0.9432

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	149	2.24	0.0267
Satterthwaite	Unequal	46.647	2.24	0.0302

Le premier tableau donne la probabilité critique bilatérale du test F d'égalité des variances. Cette probabilité étant très supérieure à 5 ou 10 % ($p\text{-}value = 0.94$), on en déduit qu'on ne peut pas rejeter l'hypothèse d'égalité des variances. Nous pouvons donc supposer que $\sigma_X = \sigma_Y$.

Le résultat du test T d'égalité des moyennes se lit donc sur la ligne *Pooled* du second tableau. Avec une probabilité critique $p\text{-}value_U = p\text{-}value_B/2 = 1.33\ \% < 5\ \%$, nous pouvons conclure, avec un risque de 5 %, que la consommation moyenne des habitants de Limoges est significativement supérieure à celle des habitants des autres communes.

Partie V (25 min, 4 points)

Dans le cadre de la certification Qualité ISO 9001, Limoges-Métropole a réalisé en 2023 une enquête pour évaluer ses services « cycle de l'eau ». Une des questions de cette enquête portait sur l'intérêt des usagers (particuliers et professionnels) concernant l'installation d'un compteur (d'eau) connecté permettant, en particulier, d'alerter en cas de fuite d'eau. Sur les 753 réponses, 381 sont favorables (à l'installation d'un compteur connecté), 369 sont défavorables et 3 sont sans avis.

1) La population Ω est toujours l'ensemble des usagers de LM. La variable X représente l'intérêt d'un usager concernant l'installation d'un compteur connecté : 1 s'il est favorable, 0 s'il ne l'est pas. La variable X suit donc une loi de Bernoulli $\mathcal{B}(1, p)$ où p représente la proportion d'usagers favorables (à l'installation d'un compteur connecté) dans la population. On a un échantillon de $n = 750$ personnes dont $k = 381$ sont favorables. Bien-sûr, on ne tient pas compte des 3 personnes sans avis ($753 - 3 = 750$)

2) La fréquence empirique $F = K/n$ où $K = \sum X_i$ est un estimateur sans biais et convergent de p. Une estimation de p est donc $f = k/n = 381/750 = 50.8\ \%$. L'intervalle de confiance à 95 % pour p est

$$\mathrm{IC}_{1-\alpha}(p) = \left[F \pm z_{0.975}\sqrt{\frac{F(1-F)}{n}}\right] \implies \mathrm{ic}_{1-\alpha}(p) = \left[0.508 \pm 1.96\sqrt{\frac{0.508(1-0.508)}{750}}\right] = [47.22\ \%, 54.38\ \%]$$

3) Pour affirmer que plus de la moitié des usagers est favorable à l'installation d'un compteur connecté, nous devons effectuer le test $H_0 : p \leqslant 0.50$ contre $H_1 : p > 0.50$. La variable de décision est (avec $p_0 = 0.50$)

$$Z = \frac{F - p_0}{\sqrt{p_0(1-p_0)/n}} \overset{\sim}{\cdot} \mathcal{N}(0,1) \implies z = \frac{0.508 - 0.50}{0.50/\sqrt{750}} = 0.438$$

En prenant un risque de 5 %, on obtient la région critique (à droite) : $W =]1.645, +\infty[$. Comme $z = 0.43 \notin W$, on ne peut par rejeter H_0 : les données collectées ne permettent pas de conclure que plus de la moitié des usagers est favorable à l'installation d'un compteur connecté.