Faculté de Droit & des Sciences Économiques — Université de Limoges

ANNÉE UNIVERSITAIRE 2022-2023

Session 1 **Semestre 4**

Licence Economie-Gestion – 2e année

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

Problème

En 2022, le poids des déchets résiduels (déchets non recyclables déposés dans les « bacs verts ») sur l'agglomération de Limoges-Métropole était en moyenne de 212 kg par an et par habitant, avec un écart-type de 35 kg. (*Source moyenne : Magazine « Le Métropol » n°85, mars-avril 2023.*) La suite du sujet est librement inspirée de ces données.

Suite à la mise en place de la collecte des bio-déchets, Limoges-Métropole a souhaité estimer l'évolution du poids moyen des déchets résiduels pendant le premier trimestre 2023. Pour cela, elle a pesé les déchets résiduels d'un échantillon de 111 personnes de janvier à mars 2023 (les données seront ramenées à des données annuelles).

Dans la suite, on suppose que le poids des déchets résiduels d'une personne suit une loi normale.

Partie I (20 min, 4 points)

Les mesures faites sur l'échantillon sont résumées dans la sortie SPSS suivante :

Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
Poids	111	117.90	304.50	202.1225	3.67818	38.75198
Valid N (listwise)	111					

1) On étudie sur la population (Ω) des habitants de l'agglomération, le poids (X) des déchets résiduels (en 2023). On suppose que X suit une loi normale $\mathcal{N}(\mu, \sigma)$ où μ représente le poids moyen des déchets résiduels et σ l'écart-type de ce poids, les deux étant inconnus (en 2023). Pour cela, on utilise un 111-échantillon $(X_1, \dots, X_{111})$ correspondant à 111 variables aléatoires indépendantes et de même loi que X.

2) Le poids des déchets de l'échantillon des $n = 111$ personnes varie de 117.9 à 304.5 kg, avec un écart-type de $s = 38.75$ kg. Le poids moyen sur l'échantillon ($\bar{x} = 202.12$ kg) est une estimation sans biais de μ. La précision de cette estimation est donnée par l'erreur standard se = 3.68.

3) Comme X suit une loi normale, on a

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \hookrightarrow \mathrm{St}(n-1) \implies \mathrm{ic}_{0.95}(\mu) = \left[\bar{x} \pm t_{0.975}\frac{s}{\sqrt{n}}\right] = [194.87, 209.41]$$

où $t_{0.975} = 1.982$ est lu dans la table St(110).

Partie II (30 min, 5 points)

On souhaite vérifier que les variations de poids des déchets n'ont pas augmenté (par rapport à 2022).

1) Il s'agit d'un test sur une variance. L'hypothèse nulle est l'hypothèse de stabilité soit $H_0 : \sigma = \sigma_{2022} = 35$. L'hypothèse alternative correspond à une augmentation des variations, soit $H_1 : \sigma > 35$.

2) La variable de décision est

$$K^2 = \frac{(n-1)S^2}{\sigma^2_{2022}} \hookrightarrow \chi^2(n-1) \implies k^2 = \frac{110 \times 38.75^2}{35^2} = 134.85$$

Avec un risque de première espèce de 5 %, la région critique (à droite) est $W =]k^2_{0.95}, +\infty[=]135.5, +\infty[$. Comme $k^2 \notin W$, on ne peut pas rejeter l'hypothèse H_0.

3) Les données collectées ne démontrent pas une augmentation des variations du poids des déchets résiduels en 2023. Le risque pris est le risque de seconde espèce $\beta = P(H_0|H_1)$ qui est inconnu (mais calculable).

4) Commenter la sortie R suivante :

```
R> VarTest(dechets$Poids, sigma.squared=35**2)
        One Sample Chi-Square test on variance
data:  dechets$Poids
X-squared = 134.85, df = 110, p-value = 0.1079
alternative hypothesis: true variance is not equal to 1225
95 percent confidence interval:
 1172.245 1993.419
sample estimates:
variance of x
     1501.716
```

La sortie R correspond à un test bilatéral sur la variance avec comme hypothèse nulle $H_0 : \sigma^2 = 35^2$. On retrouve la valeur de l'observation de la variable de décision $k^2 = 134.85$ (X-squared). La *p-value* de 10.79 % est celle associée au test bilatéral. La *p-value* correspondant à notre test (à droite) est donc *p-value* $= 10.79/2 = 5.4$ %. Comme elle supérieure au risque $\alpha = 5$ %, cela confirme qu'on ne peut pas rejeter l'hypothèse H_0. Toutefois, en prenant un risque tout juste supérieur, par exemple $\alpha = 6$ %, la conclusion aurait été inverse.

Partie III (30 min, 5 points)

On souhaite à présent tester si le poids moyen des déchets résiduels a baissé en 2023 (par rapport à 2022).

1) Il s'agit d'un test sur une moyenne. L'hypothèse nulle est l'hypothèse de stabilité soit $H_0 : \mu = \mu_{2022} = 212$. L'hypothèse alternative correspond à une diminution du poids moyen, soit $H_1 : \mu < 212$.
La variable de décision est

$$T = \frac{\overline{X} - \mu_{2022}}{S/\sqrt{n}} \hookrightarrow \mathrm{St}(n-1) \implies t = \frac{202.12 - 212}{38.75/\sqrt{111}} = -2.685$$

Avec un risque de première espèce de 5 %, la région critique (à gauche) est $W =]-\infty, t_{0.05}[=]-\infty, -1.659[$. Comme $t \in W$, on rejette l'hypothèse H_0 et on accepte l'hypothèse H_1.

2) Les données collectées démontrent une diminution du poids moyen des déchets résiduels en 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

$$p\text{-}value = P(T < -2.685) = 1 - P(T < 2.685) \approx 1 - 0.995 = 0.5\ \%$$

La probabilité critique étant inférieure au risque de première espèce ($\alpha = 5$ %), cela confirme le rejet de l'hypothèse nulle. On voit en outre que même en prenant un risque beaucoup plus faible ($\alpha = 1$ %) la conclusion aurait été la même.

4) Commenter la sortie STATA suivante :

```
Stata. ttest Poids == 212

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   Poids |     111    202.1225    3.678176    38.75198    194.8332    209.4118
------------------------------------------------------------------------------
    mean = mean(Poids)                                            t =  -2.6854
Ho: mean = 212                                 degrees of freedom =      110

   Ha: mean < 212               Ha: mean != 212                 Ha: mean > 212
 Pr(T < t) = 0.0042         Pr(|T| > |t|) = 0.0084          Pr(T > t) = 0.9958
```

La sortie STATA correspond à un test T sur une moyenne, d'hypothèse nulle $H_0 : \mu = 212$. La probabilité critique de notre test est celle se trouvant à gauche soit *p-value* $= 0.0042$, très proche de la valeur trouvée précédemment 0.005. (On remarque aussi l'intervalle de confiance à 95 % de la moyenne que nous avions calculé dans la partie I !)

Partie IV (20 min, 3 points)

Sur les 111 personnes formant l'échantillon, 36 effectuent le tri des bio-déchets (en utilisant un composteur individuel ou collectif).

1) La population étudiée correspond toujours aux habitants de l'agglomération. La variable étudiée X est une variable dichotomique indiquant si la personne effectue le tri des bio-déchets ($X(\omega) = 1$) ou non ($X(\omega) = 0$). La variable X suit donc une loi de Bernoulli $\mathcal{B}(1, p)$ où p représente la part (proportion) d'habitants effectuant le tri des bio-déchets dans l'agglomération. On note $(X_1, \ldots, X_{111})$ l'échantillon correspondant.
Comme n est suffisamment grand, on a

$$K = \sum_{i=1}^{n} X_i \qquad F = \frac{K}{n} \underset{\sim}{\hookrightarrow} \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

La fréquence sur l'échantillon $f = k/n = 36/111 = 32.43$ % est une estimation sans biais de p. Un intervalle de confiance à 95 % pour p est donné par

$$\mathrm{ic}_{0.95}(p) = \left[f \pm z_{0.975}\sqrt{\frac{f(1-f)}{n}}\right] = [23.72\ \%, 41.14\ \%]$$

2) La sortie SAS ci-dessous permet-elle de conclure que plus d'un quart de la population effectue le tri des bio-déchets ? (Vous devez, bien entendu, justifier théoriquement votre réponse.)

Biodechets				
Biodechets	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	75	67.57	75	67.57
1	36	32.43	111	100.00

Binomial Proportion	
Biodechets = 1	
Proportion	0.3243
ASE	0.0444

Test of H0: Proportion = 0.25	
ASE under H0	0.0411
Z	1.8084
One-sided Pr > Z	0.0353
Two-sided Pr > |Z|	0.0705

Pour répondre à la question, il est donc nécessaire d'effectuer un test $H_0 : p = 0.25$ contre $H_1 : p > 0.25$. Il s'agit donc d'un test unilatéral à droite, dont la probabilité critique est donnée dans le troisième tableau de la sortie SAS. On peut y lire que $p\text{-}value = 0.0353$ qui est inférieure à un risque $\alpha = 5\ \%$. On en déduit donc, qu'avec un risque de 5 %, on peut affirmer que plus de 1/4 de la population effectue le trie des bio-déchets.

Partie V (20 min, 3 points)

La sortie JAMOVI suivante précise le poids des déchets résiduels selon que la personne effectue le tri des bio-déchets ou non.

Descriptives

	Biodechets	N	Missing	Mean	Median	SD	Minimum	Maximum
Poids	Non	75	0	206	207	39.3	126	305
	Oui	36	0	194	190	36.7	118	270

1) On note X la variable aléatoire représentant le poids des déchets résiduels des personnes n'effectuant pas le tri des bio-déchets, et Y celle de ceux effectuant ce tri. Les paramètre $\mu_X = \mathrm{E}(X)$ (resp. $\mu_Y = \mathrm{E}(Y)$) représente le poids moyen des déchets résiduels des personnes n'effectuant pas le tri des bio-déchets (resp. effectuant ce tri). De même σ_X et σ_Y représentent les écarts-types de chaque des variables.
D'après la sortie JAMOVI, on possède un 75-échantillon de la première, avec un poids moyen $\bar{x} = 206$ et un 36-échantillon de la seconde avec un poids moyen $\bar{y} = 194$. La sortie précise en outre, minimum maximum et écart-type : $s_X = 39.3$ et $s_Y = 36.7$. On voit donc que sur cet échantillon, $\bar{x} > \bar{y}$.

2) Les sorties JAMOVI suivantes permettent-elles d'affirmer que les personnes effectuant le tri des bio-déchets produisent moins de déchets résiduels ?

Homogeneity of Variances Test (Levene's)

	F	df	df2	p
Poids	0.193	1	109	0.661

Note. A low p-value suggests a violation of the assumption of equal variances

Independent Samples T-Test

		Statistic	df	p		Effect Size
Poids	Student's t	1.53	109	0.064	Cohen's d	0.310

Note. $H_a\ \mu_{Non} > \mu_{Oui}$

On souhaite donc tester $H_0 : \mu_X = \mu_Y$ contre $H_1 : \mu_X > \mu_Y$. On commence par tester l'égalité (l'homogénéité) des variances $H_0 : \sigma_X = \sigma_Y$ par un test de Levene (bilatéral). La probabilité critique du test (premier tableau) étant de 66 %, on ne peut pas rejeter l'hypothèse H_0 d'égalité des variances. Cela permet de tester ensuite l'égalité des moyennes via un test T (second tableau). La probabilité critique (unilatérale) étant de 6.4 %, il n'est pas possible de rejeter l'hypothèse H_0 avec un risque de 5 %. Le poids moyen des déchets résiduels des personnes n'effectuant pas le tri des bio-déchets n'est pas significativement supérieur à celui de ceux l'effectuant, avec un risque de 5 %.
Si on s'autorise un risque supérieur, par exemple de 10 %, la conclusion s'inverse, et on conclut que $\mu_X > \mu_Y$. Toutefois, avec une taille d'effet de 0.31 (d de Cohen), la différence peut être considérée comme relativement faible.