TD N° 7 : Comparaison d'échantillons

Exercice 1

Une entreprise fabrique un médicament sur deux chaînes de production. On s'intéresse aux variations de la quantité d'une certaine substance A contenue dans chaque médicament. On a contrôlé le dosage de la substance A avec un échantillon de 100 médicaments à la sortie de chacune des deux chaînes de fabrication. On a trouvé un dosage moyen de 10.75mg pour la première chaîne et 10.70mg pour la deuxième. On sait par ailleurs que l'écart-type des chaînes de production est le même et est égal à 0.2mg.

Construire un test à 1% permettant de savoir si la différence des moyennes observées est due à des fluctuations de l'échantillonnage ou bien si la chaîne de fabrication n°1 produit des médicaments contenant davantage de substance A que la chaîne n°2.

Hypothèses : Soit X le dosage de la substance A dans un médicament de la chaine n°1
Y 2

On note $E(X)=\mu_1$, $E(Y)=\mu_2$ et $V(X)=\sigma^2=V(Y)=0.2^2$

On teste

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 > \mu_2 \end{cases} \Leftrightarrow \begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 > 0 \end{cases}$$

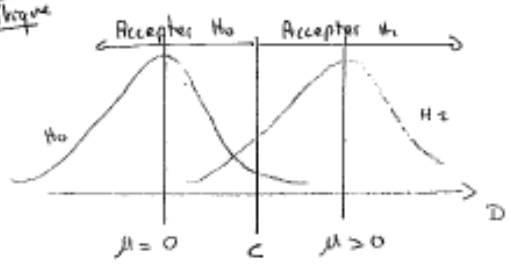
variable de décision : Les moyennes $\bar{X}$ et $\bar{Y}$ sont des estimateurs usuels de μ_1 et μ_2 donc variables de décision. L'échantillon est de grande taille donc on considère d'après le T.C.L que

$$\bar{X} \rightsquigarrow \mathcal{N}\left(\mu_1, \frac{\sigma^2}{n}\right) \text{ et } \bar{Y} \rightsquigarrow \mathcal{N}\left(\mu_2, \frac{\sigma^2}{n}\right)$$

Les variables $\bar{X}$ et $\bar{Y}$ sont indépendantes donc

$$D = \bar{X} - \bar{Y} \rightsquigarrow \mathcal{N}\left(\mu_1 - \mu_2, \frac{2\sigma^2}{n}\right)$$

Allure de la région critique : On pose $\mu = \mu_1 - \mu_2$



la région critique est

$$W = \{ D \geq c \}$$

calcul du seuil

$\alpha = P[W \mid H_0 \text{ vraie}]$

Supposons H_0 vraie alors $D \sim \mathcal{N}\left(0, \frac{2\sigma^2}{n}\right)$ d'où

$\alpha = P[D \geq c] = P\left[\sqrt{\frac{n}{2}}\frac{1}{\sigma} D \geq \sqrt{\frac{n}{2}} \times \frac{1}{\sigma} c\right]$

$\Leftrightarrow 0.01 = P[Z \geq c'] \overset{\text{table}}{\Rightarrow} c' = 2.33 \Rightarrow c = \frac{2.33 \times \sigma \times \sqrt{2}}{\sqrt{n}} = 0.066$

Décision Sur l'échantillon on obtient

$d = \bar{x} - \bar{y} = 10.75 - 10.70 = 0.05 < c$

donc on garde l'hypothèse H_0, i.e on considère que la différence observée est due à la fluctuation de l'échantillon. On ne connait pas l'erreur de se tromper en prenant cette décision puisque β est inconnu

Exercice 2

Un sondage effectué auprès de 2000 personnes indique que 19% d'entre elles connaissent la marque de lessive Omopaic. Après une campagne publicitaire, un sondage analogue auprès de 1000 personnes montre que 230 d'entre elles connaissent cette marque. Peut-on considérer que la campagne a été efficace ?

Hypothèses : Soient

- p_1 le taux de personnes dans toute la population connaissant la marque avant campagne
- p_2 le taux de personnes dans toute la population connaissant la marque après campagne

On teste les hypothèses :

c'est l'hypothèse privilégiée mais par valeurs liées donc dans H_2 ⟶ $\begin{cases} H_0 : p_2 = p_1 \\ H_2 : p_2 < p_1 \end{cases} \Leftrightarrow \begin{cases} H_0 : p_2 - p_1 = 0 \\ H_2 : p_2 - p_1 < 0 \end{cases}$

variables de décision $\bar{X}_1$ la fréquence empirique sur échantillon 1 et $\bar{X}_2$ celle de échantillon sont des estimateurs sans biais de p_1 et p_2 donc sont choisis comme variables de décision. On a des échantillons de grande taille donc le T.C.L permet de considérer que

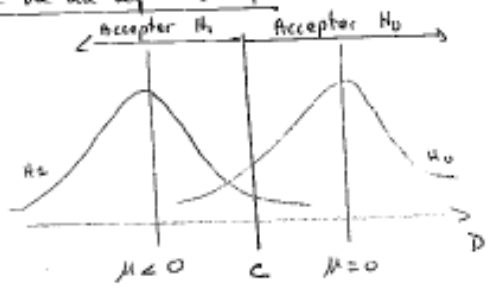
$\bar{X}_1 \sim \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n_1}\right)$ et $\bar{X}_2 \sim \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$

On note $D = \bar{X}_1 - \bar{X}_2$. Etant donné que $\bar{X}_1$ et $\bar{X}_2$ sont indépendantes

$D \sim \mathcal{N}(\mu, \sigma^2)$

où $\mu = p_1 - p_2$ et $\sigma^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$

Allure de la région critique



$W = \{ D \leq c \}$

Calcul du seuil

$\alpha = P[W \mid H_0 \text{ vraie}]$

Supposons H_0 vraie alors $D \hookrightarrow \mathcal{N}(0, \sigma^2)$ avec $p_1 = p_2 = p$ donc

$\sigma^2 = p(1-p)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]$ d'où

$\alpha = P[D \leq c] = P\left[\frac{D}{\sigma} \leq \frac{c}{\sigma}\right] = P[Z \leq c']$ où $Z \hookrightarrow \mathcal{N}(0,1)$

Si on suppose un risque $\alpha = 0.05$ on a

$c' = -1.64 \Rightarrow c = -1.64 \times \sigma$

On remplace p par une moyenne pondérée de $\bar{x}_1$ et $\bar{x}_2$ d'où

$\hat{p} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = 0.203$ (exo 6 TD n°3)

D'où $c = -0.085$

Décision

Sur l'échantillon on a

$d = \bar{x}_1 - \bar{x}_2 = 0.19 - 0.23 = -0.04 < c$

donc on accepte H_1, i.e on considère que la campagne a eu un impact avec 5% de chance de se tromper.

Exercice 3

Une enquête sur la consommation annuelle des ménages est réalisée par l'INSEE régulièrement. Ces ménages sont répartis en 5 grandes catégories suivant leur localisation :

C1 : ménages en zone rurale,

C2 : ménages résidant dans une unité urbaine inférieure à 20000 habitants,

C3 : ménages résidant dans une unité urbaine comprise entre 20000 habitants et 100000 habitants,

C4 : ménages résidant dans une unité urbaine supérieure à 100000 habitants autre que l'agglomération parisienne,

C5 : ménages résidant dans l'agglomération parisienne.

Un groupement commercial s'intéresse particulièrement à la consommation annuelle des produits contenus dans la nomenclature 17 de l'INSEE c'est-à-dire,

la consommation annuelle en mouton, agneau et chevreau et il souhaite savoir s'il y a un effet "localisation" sur la consommation annuelle moyenne des ménages pour ces produits. Le groupement commercial interroge 5 ménages par catégories. Les résultats en euro sont :

C1	C2	C3	C4	C5
56	47	55	61	69
66	50	51	62	71
54	55	59	54	55
61	46	54	54	62
56	56	59	62	53

On suppose que, pour tout $i \in \{1,...,5\}$, la consommation annuelle d'un ménage en euro de catégorie Ci peut être modélisée par une var X_i suivant la loi normale $N(\mu_i,\sigma^2)$, avec μ_i et σ inconnus.

1. On donne sceT= 908,64 et sceR= 556,40. Dresser le tableau ANOVA.

De la décomposition de la variance on déduit
sceA=sceT-sceR=908,64-556,40=352,24.
Par ailleurs on a n=5*5=25 et p=5
On obtient le tableau suivant

Source de variation	Somme des carrés	ddl	Carré moyen
Expliquée	352,24	4	88,06
Résiduelle	908,64	20	27,82
Totale	556,40	24	

2. Effectuer, au risque 5%, le test ANOVA :

H_0:μ_1=...=μ_5
contre
H_1: "il existe au moins 2 moyennes différentes"

Interpréter le résultat.

La statistique du test vaut F=88,06/27,82=3,16. Or sous l'hypothèse H_0, F suit une loi de Fisher à p-1=4 et n-p=24 ddl. Avec α=5%, on obtient un seuil C=2,87. On a F>C, donc on conclut qu'il y a un impact de la catégorie sur la consommation avec un risque de 5% de se tromper

3. Peut-on affirmer, au risque 5%, que $\mu_3 \neq \mu_4$?

On applique la méthode de Scheffé. On doit vérifier si la différence des moyennes des échantillons est supérieure ou inférieure au seuil

$$\sqrt{4 * 2,87 * 45,432} * \sqrt{\frac{1}{100} + \frac{1}{100}} = 3,23$$

Or

$$\left|\bar{x}_3 - \bar{x}_4\right| = \left|55{,}6 - 58{,}6\right| = 3 < 3{,}23$$

On conclut qu'il n'y a pas de différence significative entre les catégories C_3 et C_4.