RAPPEL DES NOTIONS DE STATISTIQUE

1. Définition d'une méthode STATISTIQUE

Une méthode statistique a pour but de tirer des conclusions (estime un paramètre ou tester une hypothèse) d'une population d'individus, plantes ou d'animaux ou de plusieurs populations à partir d'échantillons prélevés au sein de cette (ces) population(s).

Exemple 1 – Echantillon extrait d'une population de pommiers

Soit un verger de 50 pommiers, on choisit au hasard (tirage au sort à l'aide d'une table de nombres aléatoires) 4 pommiers que l'on soumet à un traitement

Ces pommiers forment un échantillon aléatoire et simple (chaque pommier du verger a la même chance d'être prélevé pour être inclus dans l'échantillon) extraits de la population « pommiers du verger ».

On observe, par exemple, le rendement en fruits sur ces 4 pommiers et on obtient; les résultats suivants :

Arbre1 30,10 kg

Arbre2 27,30 kg

Arbre3 31,30 kg

Arbre4 34,8 kg

2. Définition des indicateurs statistique

2.1 La moyenne de l'échantillon:

c'est la somme de toutes les observations divisées par le nombre d'observations.

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ Avec xi = ième observation de l'échantillon (i varie de 1 jusqu'à n)

n = nombre total d'observations.

Exemple 2 : Calcul de la moyenne de l'échantillon présenté à l'exemple 1

La moyenne de l'échantillon $\bar{x} = 30{,}87$ kg

2.2 La variance de l'éçhanti11on (et l'écart-type)'

Elle mesure la dispersion des observations de l'échantillon autour de la moyenne

Elle se calcule en prenant la somme des carrés des écarts par rapport à la moyenne ($\bar{x}$) divisée par le nombre d'observations

$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ Ou encore $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \frac{SCE}{n}$

- n : nombre total d'observations
- $(x_i - \bar{x})^2$: somme des carrés des écarts (S.C.E)
- $\frac{(\sum_{i=1}^{n} x_i)^2}{n}$: est appelé terme correctif

Exemple 3 : Calcul de la variance à partir de l'échantillon de l'exemple 1.

La somme des observations = **123,5**

La somme des carrés des écarts : **SCE**

SCE = [(30,1² +27,3² +31,3² +34,8²) - (30,1 +27,3 +31,3 +34,8)²/4] = **28,97**

La variance de l'échantillon **S²** = (28,97/4) = 7,2425 kg²

L'écart-type : racine carrée de la variance de l'échantillon

$\boldsymbol{S = \sqrt{S^2} = \sqrt{7,2425} = 2,69\ kg}$ (en moyenne, la variabilité absolue autour de la moyenne = 2,69 kg)

Le coefficient de variation = écart-type de l'échantillon exprimé en pourcentage de la moyenne.

$\boldsymbol{CV = \frac{S}{\bar{x}} \times 100}$ = 2,69/ 30,87 = 8,7%

(en moyenne, la variabilité relative autour de la moyenne est de 8,7 %)

3. Méthodes statistique relatives aux moyennes

L'Utilisation de ces méthodes suppose que certaines conditions d'application soient vérifiées :

- normalité de la ou des populations considérées
- le ou les échantillons doivent être aléatoires, simples et indépendants

Dans le cas où l'on compare les paramètres de deux ou plusieurs populations il faudra supposer, l'égalité des variances de ces populations.

3.1 Intervalle de confiance de la moyenne de la population (N < 30)

Problème Quelle est « la confiance » que l'on peut accorder à l'estimation de la moyenne de la population à partir de celle d'un échantillon ou en d'autres terme, entre quelles limites peut-on affirmer que la moyenne de la population doit se situer, avec un risque de se tromper, (risque d'erreur) de 5%

Les limites de confiance de la moyenne peuvent être déterminées de la façon :

$\bar{x} \pm t_{1-\alpha/2} \cdot \sigma_D$ σ_D : erreur (écart) commise sur l'estimation de la moyenne.

$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ Elle s'estime à partir de l'écart-type de la population et de l'effectif de l'échantillon $\bar{x} \pm t_{1-\alpha/2} \cdot \sqrt{\frac{SCE}{n(n-1)}}$

Où $t_{1-\alpha/2}$ est une quantité lue dans la table t de STUDENT dont le paramètre d'entrée (nombre de degrés de liberté) k = n – 1 permettant de préciser le degré de confiance.

Si on souhaite estimer l'intervalle avec un risque d'erreur de 5%, la valeur t est lue avec une probabilité 1-0,05/2 = 0,975

Exemple

Calculer l'intervalle de confiance de la moyenne des rendements en fruits des pommiers de l'échantillon de ***l'exemple 1*** *pour* $\alpha = 5\%$ *:*

avec $\bar{x} = 30{,}87\ kg$ *, SCE = 28,97*

on a : $\bar{x} \mp t_{(1-\frac{\alpha}{2})} \sqrt{\frac{SCE}{n(n-1)}}$ *,soit* $30{,}87 \mp t_{0,975} \sqrt{\frac{28{,}97}{4\times3}}$ *avec k=ddl= 4-1= 3*

donc $(30{,}87 \mp 4{,}94) \Rightarrow$ *c'est-à-dire la limite supérieur 25,93 et limite inferieur 35,81*

3.2 Comparaison des moyennes de 2 populations.

Problème. On veut savoir si les moyennes de 2 populations (μ_1, μ_2) sont statistiquement égales ou s'il n'y a pas de différences significatives entre les moyennes.

On suppose au départ que les conditions d'application sont remplies.

On teste l'hypothèse : $\mu_1 = \mu_2$

(μ_1 et μ_2 sont les moyennes des 2 populations)

Le résultat du test sera soit on accepte l'hypothèse

soit on rejette l'hypothèse

Le test ne peut être réalisé qu'à partir d'un échantillon prélevé au sein de chaque population.

Exemple 7 - Soit un verger de 50 pommiers, on a choisi au hasard 4 pommiers qui ont été soumis au traitement n°1 et toujours au hasard 5 pommiers qui ont été soumis au traitement n°2. Les résultats obtenus pour ces 2 traitements sont les suivantes :

TRAITEMENT N° 1	TRAITEMENT N°2
30,1 kg 27,3 kg 31,3 kg 34,8 kg	36,7 kg 39,5 kg 37,4 kg 41,2 kg 34,2 kg

Ces traitements constituent 2 populations de pommiers pour lesquels on possède un échantillon de 4 pommiers (traitement1) et 5 pommiers (traitement 2)

L'objectif du test sera de savoir s'il existe des différences significatives de rendements en pommes entre les 2 traitements.

Le test nécessite le calcul de la quantité suivante:

$$t_{obs} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{SCE_1 + SCE_2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

L'indice 1 = traitement n°1
L'indice 2 = traitement n°2
avec k = n_1+n_2-2 degrés de liberté

En théorie, cette quantité est une valeur de la distribution t de STUDENT lorsque l'Hypothèse $\mu_1 = \mu_2$ est vraie.

Pour tester l'hypothèse, il suffira de vérifier que cette quantité calculée n'est pas plus élevée que celle de la distribution théorique t de STUDENT.

Avec un risque d'erreur de 5% (soit un degré de confiance de 95), on rejettera l'hypothèse d'égalité des 2 moyennes.

Lorsque $t_{obs} \geq \begin{cases} t_{1-\alpha/2} \\ t_{0,975} \\ \text{avec } k = n_1 + n_2 - 2 \text{ degrés de liberté} \end{cases}$

Exemple 8 – Test d'égalité de moyenne des 2 traitements effectués sur pommier •

Hypothèse m1 = m2 (égalité des moyennes des populations)

On calcule le t_{obs} , en réalisant les calculs suivants:

	TRAITEMENT N° 1	TRAITEMENT N°2
1 2 3 4 5	30,1 kg 27,3 kg 31,3 kg 34,8 kg	36,7 kg 39,5 kg 37,4 kg 41,2 kg 34,2 kg
n	4	5
$\bar{x}$	30,875	37,8
SCE	28,9675	28,78

Solution

$$t_{obs} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{SCE_1 + SCE_2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{|30{,}875 - 37{,}8|}{\sqrt{\frac{28{,}9675 + 28{,}78}{4+5-2}\left(\frac{1}{4} + \frac{1}{5}\right)}} = 3{,}59$$

la valeur de $t_{(1-\alpha/2)}$ avec un risque d'erreur de 5%, et un degré de liberté de $\mathbf{n_1+n_2-2 = 4+5-2 = 7}$ lue sur la table de STUDENT $t_{0,975} = 2{,}36$

Puisque $t_{obs} < t_{(1-\alpha/2)}$ Donc en déduit qu'il existe une différence significative entre les moyennes des rendements en pommes entre les 2 traitements.