LES STATISTIQUES INFERENTIELLES

(test de Student)

L'inférence statistique est la partie des statistiques qui, contrairement à la statistique descriptive, ne se contente pas de décrire des observations, mais extrapole les constatations faites à un ensemble plus vaste et permet de tester des hypothèses sur cet ensemble ainsi que de prendre des décisions.

Un test statistique est un mécanisme qui permet de trancher entre deux hypothèses au vu des résultats d'un échantillon.

Soient H_0 et H_1 deux hypothèses (H_0 est appelée hypothèse nulle, H_1 hypothèse alternative), dont une et une seule qui est vraie. La décision consiste à retenir H_0 ou H_1.

- Pour un test bilatéral, nous pouvons émettre les hypothèses suivantes :

- Hypothèse nulle, H_0 : $p_A = p_B$
- Hypothèse alternative, H_1 : $p_A \neq p_B$.

- Pour un test unilatéral, les hypothèses deviennent :

- Hypothèse nulle, H_0 : $p_A = p_B$
- Hypothèse alternative, H_1 : $p_A > p_B$ ou $p_A < p_B$

I. LES TESTS PARAMETRIQUES

Un test est dit paramétrique si son objet est de tester une hypothèse relative à un ou plusieurs paramètres d'une variable aléatoire qui suit la loi normale ou ayant un effectif important ($n > 30$).

1. Le test de Student

Ce test permet de comparer :

- une moyenne d'un échantillon à une valeur donnée
- les moyennes de deux échantillons indépendants
- les moyennes de deux échantillons appariés.

L'emploi de ce test reste subordonné en général à deux conditions d'application importantes qui sont la normalité et le caractère aléatoire et simple des échantillons.

La première condition n'est toutefois pas essentielle lorsque les échantillons ont des effectifs suffisants (en pratique, la valeur de 30 est souvent retenue) pour assurer la quasi-normalité des distributions d'échantillonnage des moyennes. En plus, de ces deux conditions, nous devrons supposer, dans certains tests relatifs aux moyennes, l'égalité des variances des échantillons considérées.

a. Cas d'un seul échantillon

Le test de Student cas d'un seul échantillon est aussi appelé **test de conformité**, ce test a pour but de vérifier si notre échantillon provient bien d'une population avec la moyenne spécifiée, μ_0, ou s'il y a une différence significative entre la moyenne de l'échantillon et la moyenne présumée de la population.

Exemple: Une usine veut vérifier le bon fonctionnement de ces machines car l'usure des machines peut impliquer une déviation aux normes imposées.

Nous tirons aléatoirement un certain nombre d'éléments de la production, nous calculons la moyenne et nous comparons celle-ci avec la norme imposée. Les hypothèses à tester sont :

- hypothèse nulle : $H_0 : \mu = \mu_0$
- hypothèse alternative :
    - $H_1 : \mu > \mu_0$ (test unilatéral à droite)
    - $H_1 : \mu < \mu_0$ (test unilatéral à gauche)
    - $H_1 : \mu \neq \mu_0$ (test bilatéral symétrique)

Conditions d'application du test de Student : Le caractère de l'échantillon étant supposé aléatoire, l'hypothèse de normalité de la variable X doit être vérifiée (par exemple) avec le test de Kolmogorov-Smirnov si n < 30.

Calcul : Soit X une variable aléatoire distribuée selon une loi normale, la variable aléatoire définie ci-dessus suit une loi de Student avec n - 1 degrés de liberté.

$$\mathbf{t_{obs}} = \frac{\overline{X} - \mu_0}{\sqrt{\frac{S^2}{n}}}$$

où μ_0 est la moyenne de la population spécifiée par H_0, $\bar{X}$ est la moyenne de l'échantillon, S² est la variance de l'échantillon et n est la taille de l'échantillon

On compare la valeur calculée de t (t_{obs}) avec la valeur critique appropriée de t avec n - 1 degrés de liberté. On rejette H_0 si la valeur absolue de t_{obs} est supérieure à cette valeur critique.

Les valeurs critiques pour différents degrés de liberté et différents seuils de signification sont données par la table de Student. Pour un test unilatéral, nous prendrons la valeur $t_{n-1,1-\alpha}$ de la table et pour un test bilatéral, nous prendrons $t_{n-1,1-\alpha/2}$.

b. **Cas de deux échantillons indépendants**

Etant donné deux échantillons de taille n_1 et n_2, on admet qu'ils ont été prélevés d'une même population relativement à la variable étudiée, ces deux échantillons ayant été prélevés indépendamment l'un de l'autre ?

Les hypothèses à tester sont :

- hypothèse nulle : $H_0 : \mu_1 = \mu_2$

- hypothèse alternative qui prend trois formes :
    - $H_1 : \mu_1 > \mu_2$ (test unilatéral à droite)
    - $H_1 : \mu_1 < \mu_2$ (test unilatéral à gauche)
    - $H_1 : \mu_1 \neq \mu_2$ (test bilatéral)

Conditions d'application :

- Les deux échantillons sont indépendants entre eux, sont aléatoires et ont n_1 et n_2 unités indépendantes (cette condition est d'ordinaire satisfaite en utilisant une procédure de randomisation ; procédure pour laquelle on affecte au hasard chaque individu à un groupe expérimental).
- La variable aléatoire suit une loi normale ou elle a des effectifs supérieurs à 30.
- Il est aussi nécessaire de **vérifier l'égalité des variances** des échantillons (grâce au test de Fisher). **Cette condition est indispensable pour des effectifs inégaux.**

Remarques:
Plusieurs auteurs ont montré que **l'hypothèse de normalité est d'importance relativement secondaire dans le test d'égalité de deux moyennes**. En effet, dans certaines limites, la non-normalité des populations ne modifie pas sensiblement les risques d'erreur de première et deuxième espèce. Ceci est vrai surtout pour les distributions symétriques, même très différentes des distributions normales. De même, **l'hypothèse d'égalité des variances n'est**

<u>pas fondamentale au point de vue pratique lorsque les effectifs des échantillons sont égaux.</u> En raison de cette faible sensibilité du test à la non-normalité et à l'inégalité des variances, on dira qu'il s'agit, pour des effectifs égaux, **d'un test robuste**. Par contre, **<u>lorsque les effectifs des échantillons sont inégaux, il est absolument indispensable de s'assurer de l'égalité des variances</u>** et, si cette hypothèse n'est pas vérifiée, il est indispensable d'utiliser une méthode adaptée à ces circonstances. <u>On peut notamment procéder à une transformation de variable</u>, <u>destinée à stabiliser les variances, et utiliser ensuite le test de Student</u>. Cependant, ce cas **d'inégalité des variances est assez rare.**

<u>Mode de calcul :</u> On calcule la valeur t observé (t_{obs}) qui suit une variable aléatoire de Student aux degrés de liberté (ddl = n1 + n2-2).

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}}$$

où $\bar{x}_1$ *et* $\bar{x}_2$ *sont les moyennes des deux échantillons,* S_p^2 *la variance commune.* Cette dernière statistique correspond à la variance S^2 de la population parentale. Elle est égale à :

$$Vc = \frac{(n-1)*V1 + (n-1)*V2}{n1 + n2 - 2}$$

avec

$$\text{Var} = \frac{1}{n-1} * \left(\sum_{i=1}^{i=n} xi^2 - \left(\sum_{i=1}^{i=n} xi\right)^2 / n \right)$$

Ce qui revient à : $S_p^2 = \frac{SCE_1 + SCE_2}{(n_1 - 1) + (n_2 - 1)} = \frac{SCE_1 + SCE_2}{\nu_1 + \nu_2}$

Si les effectifs des échantillons sont égaux, la valeur de t devient :

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{2S_p^2}{n}}}$$

La valeur de t est comparée à la valeur critique appropriée de t (dans la table de Student) avec **($n_1 + n_2$ - 2) degrés de liberté**. On rejette H_0 si la valeur absolue de t_{obs} est supérieure à cette valeur critique. Si le test est unilatéral, nous prendrons la valeur $t_{n1 + n2 - 2, 1-\square}$ (ou $t\square_{(2),\ \square}$) de la table de Student. S'il est bilatéral, nous prendrons la valeur $t_{n1+n2-2, 1-\square/2}$ (ou $t\square_{(1),\ \square}$).

c. Cas de deux échantillons appariés:

Le test de Student pour observations pairées sert à comparer les moyennes de deux populations, dont chaque élément de l'une des populations est mis en relation avec un élément de l'autre. Par exemple, il peut s'agir de comparer deux traitements, les données étant considérées comme des paires d'observations (première observation de la paire recevant le traitement 1 et deuxième observation recevant le traitement 2).

Aspects mathématiques :

Soit x_{ij} l'observation *j* pour la paire *i* (j = 1,2 et i = 1,2,...,n). Pour chaque paire d'observations on calcule la différence $d_i = x_{i2} - x_{i1}$. Le test statistique est défini par :

$$t_{obs} = \frac{\bar{d}}{\sqrt{\frac{S_d^2}{n}}}$$

où n est le nombre de paires d'observations, $\bar{d}$ *est la moyenne des différences entre les observations et* S_d^2 *la variance.*

Le test de Student pour observations pairées est un test bilatéral. Les hypothèses sont :

- $H_0 : \mu_1 - \mu_2 = 0$ (il n'y a pas de différence entre les traitements)
- $H_1 : \mu_1 - \mu_2 \neq 0$ (il y a une différence entre les traitements)

On rejette l'hypothèse nulle au seuil de signification α si : $|t_{obs}| > t_{n-1,1-\alpha/2}$ où $t_{n-1,1-\alpha/2}$ est la valeur de la table de Student avec n - 1 degrés de liberté.

Conditions d'application :

- les échantillons ont été tirés aléatoirement
- la population des différences doit suivre une loi de Gauss. Cette condition est moins restrictive que celle de normalité des deux populations.