Inférence statistique : Tests d'hypothèse (Tests paramétriques).

Introduction

Les tests statistiques sont des méthodes de la statistique inférentielle qui permettent d'analyser des données obtenues par tirages au hasard. Ils consistent à généraliser les propriétés constatées sur des observations à la population d'où ces dernières sont extraites, et à répondre à des questions concernant par exemple la nature d'une loi de probabilité, la valeur d'un paramètre ou l'indépendance de deux variables aléatoires.

I. Tests de conformité

Les tests de conformité sont dits « *tests à 1 échantillon* ». Ils ont pour but de vérifier si un échantillon peut être considéré comme représentatif de la population dont il est extrait.

On étudie une variable quantitative X et on cherche à établir si les observations sont en accord avec la loi théorique de cette variable.

En général, il s'agit de tester si un paramètre (tel que la moyenne, la fréquence ou la variance) calculé dans l'échantillon est conforme à sa valeur au niveau de la population.

1) Tests de conformité pour une moyenne

Dans un test sur la moyenne, on prendra la moyenne empirique $\bar{X}$ comme estimateur et on posera :

$H_0: \mu = \mu_0$ Hypothèse nulle contre $H_1: \mu \neq \mu_0$ Hypothèse alternative

Dans tous les tests qui seront étudiés on supposera que les populations sont *gaussiennes*, ce que signifie que le caractère observé X peut être considéré comme une variable aléatoire suivant une loi normale $N(m, \sigma)$

a) Petit échantillon

Si $n < 30$

$$T = \frac{\bar{X} - m}{s/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - m)}{s}$$

On montre, sous l'hypothèseH_0, que la statistique T suit une loi de Student à $n - 1$ degrés de liberté :

$$T \sim t(n-1)$$

On cherche les bornes de l'intervalle d'acceptation dans une table de la loi de Student. Par exemple, dans le cas d'un test bilatéral au seuil 5%, la borne est le quantile μ_α tel que :

$$P(-\mu_\alpha \leq T \leq \mu_\alpha) = 0.95$$

Exemple

Un fabricant de diapason veut faire un contrôle de production et prélève au hasard un lot de 10 diapasons afin de mesurer la fréquence de la note qu'ils émettent. Il obtient les résultats suivants (exprimés en hertz)

428.74	426.64	438.30	439.87	440.76
436.84	450.95	426.80	442.38	431.95

Tester au seuil de 5% si la fréquence moyenne observée est conforme à la fréquence attendue de 440 Hz.

On calcule la moyenne empirique et l'écart-type empirique de l'échantillon on trouve :

$\bar{X} = 436.32 \;\; s = 7.804$

On en déduit la valeur de la statistique T :

$$T = \frac{\bar{X} - m}{s/\sqrt{n}} = \frac{436.32 - 440}{7.804/\sqrt{10}} = -1.4899$$

Le nombre de degrés de liberté est$v = n - 1 = 10 - 1 = 9$. La table de la loi de Student au risque 5% donne un quantile$\mu_\alpha = 2.26$. On ne peut donc pas rejeter l'hypothèseH_0.

b) Grand échantillon

Lorsque n est suffisamment grand, la densité de la loi de Student se rapproche remarquablement de celle de la loi normale$N(0,1)$. On considère que l'approximation est suffisante dès que ≥ 30 .

On utilise donc dans ce cas la même statistique $U = \frac{\bar{X} - m}{s/\sqrt{n}}$ mais on calcule les quantiles au moyen d'une table de la loi normale.

Exemple

L'article L3322-3 du *code de la santé publique* stipule que sont interdites, en France la fabrication et la vente de boissons apéritives à base de vin titrant plus de 18 degrés d'alcool. Les douanes ont saisi une cargaison d'importation d'un apéritif dont l'étiquette indique un titrage de 18 degrés et ont procédé à des vérifications sue 50 bouteilles : le titrage moyen observé est de 18.4 degrés avec un écart-type de 1.7 degrés.

Peut-on, au seuil 5%, considérer que ce lot respecte la législation ?

On va cette fois faire un test unilatéral. Les hypothèses du test seront :

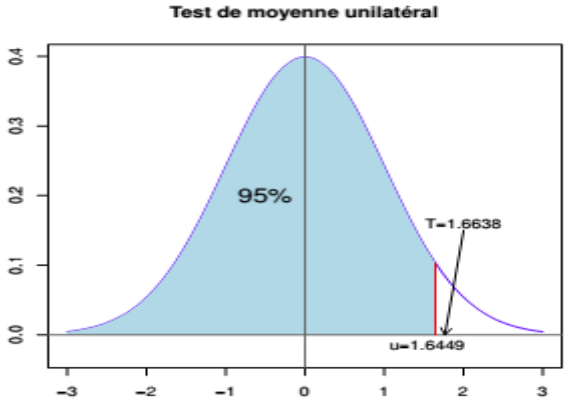
$$\begin{cases} H_0 : m = 18 \\ H_1 : m > 18 \end{cases}$$

On calcule la valeur de la statistique T :

$$U = \frac{\bar{X} - m}{s/\sqrt{n}} = \frac{18.4 - 18}{1.7/\sqrt{50}} = 1.6638$$

L'échantillon est de taille 50 est considéré comme un grand échantillon.

La table de la loi normale indique un quantile unilatéral de 1.6449 pour la probabilité de 95%. Comme$1.6638 > 1.6449$, la statistique calculée se trouve dans la région de rejet. On rejette l'hypothèse H_0 et on considère, avec un risque de 5% de se tromper, que le lot saisi présente un taux supérieur à celui autorisé.



II. Tests d'homogénéité

1) Comparaison de deux moyennes

Soit X_1 et X_2 des variables aléatoires indépendantes représentant le caractère dans la population. On suppose que X_1 et X_2 suivent une loi normal de moyennes respectivement μ_1 et μ_2, de variance respectives σ_1^2 et σ_2^2.

Si $n_1 < 30$ et $n_2 < 30$

H_0: "$\mu_1 = \mu_2$" **contre** H_1: "$\mu_1 \neq \mu_2$

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sigma \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T(n_1 + n_2 - 2)$$

Avec :

$$\sigma = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Si $T > T_{th}$ on rejette H_0

$T < T_{th}$ On accepte H_0

Exemple

On a les données suivantes

$$\alpha = 0.05 \qquad n_1 = 9 \qquad n_2 = 8$$

variable	1	2	3	4	5	6	7	8	9
X_1	21.18	20.01	22.50	22.97	21.83	23.42	18.61	25.20	22.07
X_2	22.39	21.26	22.17	25.00	22.21	20.51	22.36	24.49	

Comparer μ_1 et μ_2 ?

On calcule X_1 et X_2 et $\bar{X}_1 - \bar{X}_2$

$$X_1 = 22$$
$$X_2 = 22.5$$
$$\bar{X}_1 - \bar{X}_2 = 0.5$$
$$\sigma = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$S_1^2 = \frac{\sum(X_i - \bar{X})^2}{n - 1} = 3.7$$
$$S_2^2 = \frac{\sum(X_i - \bar{X})^2}{n - 1} = 2.28$$
$$\sigma = \sqrt{\frac{(9 - 1)3.7 + (8 - 1)2.28}{9 + 8 - 2}} = 1.7$$
$$T = \frac{22 - 22.5}{1.7 \times \sqrt{\frac{1}{9} + \frac{1}{8}}} = -0.61$$

$$H_1: m_1 \neq m_2$$

$\alpha = 0.05$ $n_1 + n_2 - 2 = 15$ $t_{0.05} = 2.131$

$T < t_{0.05}$ On accepte H_0

Si $n_1 \geq 30$ et $n_2 \geq 30$

La variable aléatoire $U = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$

$U > u_\alpha$ On rejette H_0

$U < u_\alpha$ On accepte H_0

Exemple

On veut comparer, entre les deuxUP (UP_1, UP_2), les résultats à l'examen de statistiques. On prélève pour cela 40 copies dans chaque UP. On observe les résultats suivants :

$$\bar{X}_1 = 9.4 \qquad S_1 = 2.2$$

$$\bar{X}_2 = 10.4 \qquad S_2 = 2.8$$

On exécutera le test au seuil 5%.

Il s'agit d'un test de comparaison de moyennes pour deux grands échantillons. On calcule :

$$U = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{9.4 - 10.4}{\sqrt{\frac{2.2^2}{40} + \frac{2.8^2}{40}}} = -1.776$$

$$\boldsymbol{u_\alpha = 1.96}$$

Donc $U < u_\alpha$ On accepte H_0

III. Test d'indépendance : Test de *Khi – Deux*

Il concerne deux variables aléatoires discrètes ou continues avec un nombre fini de classes et opère sur la table de contingence qui donne les effectifs croisés des deux variables. Ce test permet de tester si les deux variables peuvent être considérées comme indépendantes.

Exemple

La table suivante représente les résultats d'une enquête portant sur 300 étudiants à qui il a été demandé s'ils avaient une activité sportive régulière (S/NS) et s'ils fumaient (F/NF) :

	F	**NF**	**Total**
S	60	76	136
NS	56	108	164
Total	116	184	300

L'hypothèse H_0 est qu'il y a indépendance entre les deux variables. On va calculer, sous l'hypothèse H_0, les valeurs théoriques du tableau de contingence. On note traditionnellement $n_{i\bullet}$ les sommes en lignes et $n_{\bullet \mathrm{j}}$ les sommes en colonnes. On montre que, s'il y a indépendance parfaite entre les deux variables, les effectifs théoriques dans chaque case du tableau de contingence valent

$$\boldsymbol{n_{ij} = \frac{n_{i\bullet} \times n_{\bullet \mathrm{j}}}{N}}$$

On calcule alors la distance du χ^2 entre le tableau des valeurs observées O_{ij} et celui des valeurs théoriques n_{ij} :

$$\boldsymbol{Y = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(O_{ij} - n_{ij}\right)^2}{n_{ij}}}$$

où p est le nombre de classes de la première variable (ici $p = 2$) et q le nombre de classes de la deuxième variable (ici $q = 2$). Sous l'hypothèse H_0, la statistique Y suit une loi du χ^2 à (p − 1)(q − 1) degrés de liberté. On doit supposer aussi que l'effectif de chaque case est ≥ 5.

Reprenons l'exemple précédent et calculons le tableau d'effectifs théoriques. On trouve, en appliquant la formule (1) :

	F	**NF**
S	25.59	83.41
NS	63.41	100.59

Par exemple, la valeur de la case en haut à gauche est obtenue comme ceci:

$$\frac{116 \times 136}{300} = 52.59$$

On calcule maintenant la statistique Y

$$Y = \frac{(60 - 52.59)^2}{52.59} + \frac{(56 - 63.41)^2}{63.41} + \frac{(76 - 83.41)^2}{83.41} + \frac{(108 - 100.59)^2}{100.59} = 3.117$$

Le nombre de degrés de liberté est ici $\nu = (p-1)(q-1) = (2-1)(2-1) = 1$. La table du χ^2, indique une valeur critique égale à 3.841 au seuil 5%. Comme 3.117 < 3.841, on ne peut pas rejeter l'hypothèse d'indépendance.

- **Remarque**

Si on souhaite travailler en proportions plutôt qu'en effectifs, on pose $f_{ij} = O_{ij}/N$ (où N est l'effectif total de l'échantillon) et on note p_{ij} les probabilités théoriques. La statistique du χ^2 devient alors :

$$Y = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(Nf_{ij} - Np_{ij}\right)^2}{Np_{ij}} = N \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(f_{ij} - p_{ij}\right)^2}{p_{ij}}$$