Outils Statistiques pour la data intelligence (Module 2)

Introduction

On étudie un caractère (discret ou continu) sur une population. Un paramètre (moy., prop., var.) est inconnu ou la loi de proba de ce caractère est inconnue.

On formule ensuite des hypothèses statistiques sur la valeur ou la loi de distribution de ce paramètre.

On posera ensuite un jugement sur cette hypothèse sur base du résultat(s) obtenus sur un échantillon. Il faut que la différence entre les résultats théoriques et observées ne soient pas trop grands. La méthode qui consiste à vérifier que la différence entre ces 2 résultats n'est pas trop grande est appelée le **test d'hypothèse**.

Hypothèse Nulle, Hypothèse alternative, seuil de signification

Hypothèses

L'**hypothèse nulle** (H_0) est l'hypothèse à tester. Elle s'exprime sous la forme d'une **égalité** (=) pour les tests de valeur et d'égalité et sous la forme d'une affirmation positive pour les tests du khi-deux.

L'**hypothèse alternative** (H_1) est celle que nous accepterons si nous sommes amenées à rejeter H_0. Elle s'exprime sous la forme :

- d'une ($\neq$) $\rightarrow$ **tests bilatéraux**
- d'une (<,>) $\rightarrow$ **tests unilatéraux**
- d'une affirmation négative pour les tests du khi-deux

> Il existe des hypothèses nulles qui utilisent $\leq$ et $\geq$. L'hypothèse alternative est alors automatiquement l'inverse de l'hypothèse nulle

Test Unilatéral Vs Test Bilatéral

Le choix entre un **test unilatéral** ou **bilatéral** dépends souvent de qui demande le test. En effet, passer de l'un à l'autre peut avoir pour effet de changer la conclusion du test

Seuil de signification

Le **seuil de signification** est la probabilité de rejeter H_0 (H_1 est vrai à tord) alors que H_0 est vrai. On appelle cette erreur **erreur ou risque de** 1^{ere} **espèce**. noté α

> note : le risque de 2e espèce ou β serait le risque d'accepter H_0 alors qu'il est faux.

Démarche

1. Fixer les hypothèses H_0 et H_1
2. Préciser le seuil de signification de α

3. Préciser la loi de probabilité utilisée
4. Déterminer le seuil de rejet, la zone de non refus
5. Calculer la grandeur expérimentale
6. Tirer des conclusions

Il est important de bien poser l'hypothèse alternative !!

> car dans un test bilatéral, la zone de rejet est divisée en deux et dans un test unilatéral, la zone est concentrée d'un seul côté

Tests de valeur

Dans un **test de valeur**, on teste l'hypothèse selon laquelle le paramètre étudié de la population est égale à une valeur déterminée. L'hypothèse alternative sera alors que le paramètre étudié est différent($\neq$, $<$, $>$) de la valeur déterminée.

Existe-t-il une différence entre le résultat évalué initialement pour l'ensemble de la population et celui évalué plus tard. **Cette différence est-elle suffisamment significative pour affirmer que le paramètre de la population a changé ?**

Test de valeur d'une proportion

> rappel : dans le cas de grands échantillons ($n \geq 30$): $f \approx N\big(E(f), \sqrt{var(f)}\big)$
>
> Zone de non refus ($\alpha = 5\%$): $Pr\{f - 1.96.\sqrt{\frac{p.(p-1)}{n}} \leq p \leq f + 1.96.\sqrt{\frac{p.(p-1)}{n}}\}$
>
> Dans le cas d'un test unilattéral, un des deux côtés de cette équation n'est plus d'application

Test de valeur d'une moyenne

Echantillons de grande taille ($n \geq 30$)

($n \geq 30$): $\overline{x} \approx N\big(E(\overline{x}), \sqrt{var(\overline{x})}\big)$

et donc $\frac{\overline{x} - E(\overline{x})}{\sqrt{\overline{x}}} \approx X_G^*$

Echantillons de petite taille ($n < 30$)

($n < 30$): $\overline{x} \approx N\big(E(\overline{x}), \sqrt{var(\overline{x})}\big) \approx t_{n-1}$

Tests d'égalité

Un test d'égalité essaye de voir si il y a des similarités pour le caractère observé entre les **2 populations**

Test d'égalité de deux proportions

Nous avons 2 populations : $P_1(N_1, p_1)$et $P_2(N_2, p_2)$.

On va prendre un échantillon dans chaque population taille : n_1 et n_2 et fréquence : f_1 et f_2

On ne traite ici que des échantillons de grande taille ($n_1 \; et \; n_2 \geq 30$).

on comprends la variable d qui représente $d = f_1 - f_2$

- $\frac{d-E(d)}{\sqrt{var(d)}} \approx X_G^*$
- $E(d) = E(f_1 - f_2) = E(f_1) - E(f_2) = p_1 - p_2$
- $var(d) = var(f_1 - f_2) = var(f_1) + var(f_2) = \frac{p_1.(1-p_1)}{n_1} + \frac{p_2.(1-p_2)}{n_2}$

La P-Valeur

La p-Valeur est le plus petit niveau d'incertitude(α) en dessous duquel les données observées indiquent que l'hypothèse nulle doit être rejetée

Test d'égalité de deux moyennes

Grands échantillons ($n_1 \geq 30$ et $n_2 \geq 30$)

Nous avons 2 populations : $P_1(N_1, m_1)$et $P_2(N_2, m_2)$.

On va prendre un échantillon dans chaque population taille : n_1 et n_2 et fréquence : $\overline{x}_1$ et $\overline{x}_2$

on comprends la variable d qui représente $d = \overline{x}_1 - \overline{x}_2$

- $\frac{d-E(d)}{\sqrt{var(d)}} \approx X_G^*$
- $E(d) = E(\overline{x}_1 - \overline{x}_2) = E(\overline{x}_1) - E(\overline{x}_2) = m_1 - m_2$
- $var(d) = var(\overline{x}_1 - \overline{x}_2) = var(\overline{x}_1) + var(\overline{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

Si σ_1^2 ou σ_2^2 sont inconnus, ils seront remplacés par leur estimation S^2_{1,n_1-1} et S^2_{2,n_2-1}

Petits échantillons ($n_1 < 30$ et/ou $n_2 < 30$)

Nous avons 2 populations : $P_1(N_1, m_1)$et $P_2(N_2, m_2)$.

On va prendre un échantillon dans chaque population taille : n_1 et n_2 et fréquence : $\overline{x}_1$ et $\overline{x}_2$

on comprends la variable d qui représente $d = \overline{x}_1 - \overline{x}_2$

- $\frac{d-E(d)}{\sqrt{var(d)}} \approx t_v$ ou $v = n_1 + n_2 - 2$
- $E(d) = E(\overline{x}_1 - \overline{x}_2) = E(\overline{x}_1) - E(\overline{x}_2) = m_1 - m_2$
- $var(d) = var(\overline{x}_1 - \overline{x}_2) = var(\overline{x}_1) + var(\overline{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})$

  Si σ^2 est inconnu, il sera remplacé par :

$$\sigma^2 \approx \frac{\sum_{i_1=1}^{n_1}(x_{i_1} - \overline{x_1})^2 + \sum_{i_2=1}^{n_2}(x_{i_2} - \overline{x_2})^2}{n_1 + n_2 - 2}$$

Tests du khi-carré

Le test du Khi-deux permet de vérifier la quantité de l'ajustement entre une distribution théorique et une distribution expérimentale ou encore de tester l'indépendance entre 2 variables

Test D'ajustement

> Ex: ce dé est-il équilibré

Soit un échantillon de taille suffisamment grande sur lequel on observe les différentes modalités prises par une variable (et si besoin est, regrouper les données par classes).