Probabilités et observations économiques

C. Tuleau-Malot

Université de Nice - Sophia Antipolis

Plan

Exemples de test

Il existe différents tests :

- Tests d'hypothèse : on va comparer la moyenne sur la population à une valeur de référence
- Tests de comparaison : on va comparer la moyenne de deux populations
- Test d'adéquation : on va regarder si l'on peut considérer que notre variable suit une distribution pré-définie

Exemples de test

Il existe différents tests :

- Tests d'hypothèse : on va comparer la moyenne sur la population à une valeur de référence
- Tests de comparaison : on va comparer la moyenne de deux populations
- Test d'adéquation : on va regarder si l'on peut considérer que notre variable suit une distribution pré-définie
- Test d'indépendance : on va voir si deux variables aléatoires sont indépendantes ou non

Stratégie

Dans chacun des cas cités précédemment, on va :

1. Définir les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_a$

Stratégie

Dans chacun des cas cités précédemment, on va :

1. Définir les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_a$
2. on va calculer une statistique de test (une variable aléatoire) que l'on peut entièrement évaluer à partir d'observations et dont on connaît la loi sous l'hypothèse $\mathcal{H}_0$

Stratégie

Dans chacun des cas cités précédemment, on va :

1. Définir les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_a$
2. on va calculer une statistique de test (une variable aléatoire) que l'on peut entièrement évaluer à partir d'observations et dont on connaît la loi sous l'hypothèse $\mathcal{H}_0$
3. si la statistique de test, calculée sur l'échantillon, prend une valeur trop improbable, alors on rejette $\mathcal{H}_0$

Stratégie (2)

- $\mathcal{H}_0$ et $\mathcal{H}_a$ ne jouent pas le même rôle
- $\mathcal{H}_0$: hypothèse à laquelle on souhaite accorder le bénéfice du doute
- on rejette $\mathcal{H}_0$ que si les observations penchent fortement en faveur de $\mathcal{H}_a$

Stratégie (2)

- $\mathcal{H}_0$ et $\mathcal{H}_a$ ne jouent pas le même rôle
- $\mathcal{H}_0$: hypothèse à laquelle on souhaite accorder le bénéfice du doute
- on rejette $\mathcal{H}_0$ que si les observations penchent fortement en faveur de $\mathcal{H}_a$

Stratégie (3)

Exemple :

Une entreprise cherche à introduire un nouveau produit A, et le fait tester à 150 acheteurs potentiels, en le comparant à un produit concurrent B. 80 disent préférer le produit A. La question qui se pose est : "peut-on conclure que plus de la moitié des consommateurs préfèrent le produit A".

L'entreprise désirant montrer que son produit est meilleur, il faut prendre :
$\mathcal{H}_0$: "moins de la moitié des consommateurs préfèrent le produit A"

$\mathcal{H}_a$ est en général l'hypothèse que l'on veut démontrer.

Cadre

Dans un premier temps, nous nous limitons à des tests de comparaison de la moyenne d'une population ou d'une proportion dans une population, en faisant une comparaison à une valeur de référence.

Cadre

Dans un premier temps, nous nous limitons à des tests de comparaison de la moyenne d'une population ou d'une proportion dans une population, en faisant une comparaison à une valeur de référence.

Si on note μ la moyenne d'une population, il existe différents tests :

- $\mathcal{H}_0 : \mu \leq \mu_0$ et $\mathcal{H}_a : \mu > \mu_0$ (test unilatéral supérieur)
- $\mathcal{H}_0 : \mu \geq \mu_0$ et $\mathcal{H}_a : \mu < \mu_0$ (test unilatéral inférieur)
- $\mathcal{H}_0 : \mu = \mu_0$ et $\mathcal{H}_a : \mu \neq \mu_0$ (test bilatéral)

Cadre (2)

Deux décisions possibles :

- on ne rejette pas l'hypothèse $\mathcal{H}_0$
- on rejette l'hypothèse $\mathcal{H}_0$

Cadre (2)

Deux décisions possibles :

- on ne rejette pas l'hypothèse $\mathcal{H}_0$
- on rejette l'hypothèse $\mathcal{H}_0$

Cela revient à définir une région de rejet, on savoir un sous-espace de $\mathbb{R}^n$ où l'on décide $\mathcal{H}_a$ Pb : quand on prend une décision, on

peut se tromper. Deux sources d'erreur :

- on rejette $\mathcal{H}_0$ alors que $\mathcal{H}_0$ est vrai : erreur de première espèce
- on ne rejette pas $\mathcal{H}_0$ alors que $\mathcal{H}_a$ est vrai : erreur de deuxième espèce

Cadre (3)

Exemple :

- erreur de première espèce : conclure que plus de la moitié des consommateurs préfèrent le produit A alors que ce n'est pas vrai
- erreur de deuxième espèce : conclure que moins de la moitié des consommateurs préfèrent le produit A alors que ce n'est pas vrai

définition seuil

La probabilité de faire une erreur de première espèce est appelée seuil de signification du test

ce seuil est généralement noté α

ce seuil est généralement fixé arbitrairement au début du test

on ne contrôle pas la probabilité de la seconde erreur!

Comparaison de la moyenne à une norme

Deux cadres :

- σ^2 connu
- σ^2 inconnu

Comparaison de la moyenne à une norme : σ^2 connu

Test bilatéral :

- $\mathcal{H}_0 : \mu = \mu_0$
- $\mathcal{H}_a : \mu \neq \mu_0$

Exemple :
Une chaîne de production remplit des barils de lessive de 5kg. On pèse précisément 36 barils pour savoir si la chaîne est bien réglée, à savoir si la moyenne de remplissage est bien $\mu = 5kg$.

Pour réaliser le test, on suppose que la variable aléatoire X qui n'est autre que le remplissage du baril en kg, suit une loi $\mathcal{N}(\mu; \sigma = 0.05)$. D'après l'énoncé, on a $n = 36$.

Comparaison de la moyenne à une norme : σ^2 connu (2)

Principe :

- On ne connait pas μ mais on peut l'estimer

Comparaison de la moyenne à une norme : σ^2 connu (2)

Principe :

- On ne connait pas μ mais on peut l'estimer
- estimateur : $\bar{X}_n$

Comparaison de la moyenne à une norme : σ^2 connu (2)

Principe :

- On ne connait pas μ mais on peut l'estimer
- estimateur : $\bar{X}_n$
- estimation : on calcule $\bar{x}_n$ à partir des 36 observations et si $\bar{x}_n$ est trop éloigné de μ_0 on rejette $\mathcal{H}_0$

Comparaison de la moyenne à une norme : σ^2 connu (2)

Principe :

- On ne connait pas μ mais on peut l'estimer
- estimateur : $\bar{X}_n$
- estimation : on calcule $\bar{x}_n$ à partir des 36 observations et si $\bar{x}_n$ est trop éloigné de μ_0 on rejette $\mathcal{H}_0$
- que veut dire trop loin?

Comparaison de la moyenne à une norme : σ^2 connu (2)

- Si l'on suppose $\mathcal{H}_0$, $\bar{X}_n$ est une loi $\mathcal{N}(\mu_0; \sigma = 0.05/\sqrt{n})$

Comparaison de la moyenne à une norme : σ^2 connu (2)

- Si l'on suppose $\mathcal{H}_0$, $\bar{X}_n$ est une loi $\mathcal{N}(\mu_0; \sigma = 0.05/\sqrt{n})$
- si la valeur calculée $\bar{x}_n$ est trop improbable au vu de la distribution de $\bar{X}_n$, on rejette $\mathcal{H}_0$

Comparaison de la moyenne à une norme : σ^2 connu (2)

- Si l'on suppose $\mathcal{H}_0$, $\bar{X}_n$ est une loi $\mathcal{N}(\mu_0; \sigma = 0.05/\sqrt{n})$
- si la valeur calculée $\bar{x}_n$ est trop improbable au vu de la distribution de $\bar{X}_n$, on rejette $\mathcal{H}_0$
- sous $\mathcal{H}_0$, $\bar{X}_n \sim \mathcal{N}(\mu_0; \sigma = 0.05/\sqrt{n})$ soit $Z_n = \sqrt{n}\frac{\bar{X}_n - \mu_0}{\sigma} \sim \mathcal{N}(0; 1)$

Comparaison de la moyenne à une norme : σ^2 connu (2)

- Si l'on suppose $\mathcal{H}_0$, $\bar{X}_n$ est une loi $\mathcal{N}(\mu_0; \sigma = 0.05/\sqrt{n})$
- si la valeur calculée $\bar{x}_n$ est trop improbable au vu de la distribution de $\bar{X}_n$, on rejette $\mathcal{H}_0$
- sous $\mathcal{H}_0$, $\bar{X}_n \sim \mathcal{N}(\mu_0; \sigma = 0.05/\sqrt{n})$ soit $Z_n = \sqrt{n}\frac{\bar{X}_n - \mu_0}{\sigma} \sim \mathcal{N}(0; 1)$
- Z_n statistique de test et z_n statistique de test observée (plus $|z_n|$ loin de 0 et plus on préfère $\mathcal{H}_a$)

Comparaison de la moyenne à une norme : σ^2 connu (2)

Méthode 1 :

- On réfléchit à la région de rejet :

Comparaison de la moyenne à une norme : σ^2 connu (2)

Méthode 1 :

- On réfléchit à la région de rejet : $z_n \in]-\infty; -t[\cup]t; +\infty[$

Comparaison de la moyenne à une norme : σ^2 connu (2)

Méthode 1 :

- On réfléchit à la région de rejet : $z_n \in]-\infty; -t[\cup]t; +\infty[$
- Il faut trouver t de sorte de contrôler l'erreur de première espèce

Comparaison de la moyenne à une norme : σ^2 connu (2)

Méthode 1 :

- On réfléchit à la région de rejet : $z_n \in]-\infty; -t[\cup]t; +\infty[$
- Il faut trouver t de sorte de contrôler l'erreur de première espèce
- t est le fractile d'ordre $1 - \alpha/2$ d'une loi normale

Comparaison de la moyenne à une norme : σ^2 connu (2)

Méthode 1 :

- On réfléchit à la région de rejet : $Z_n \in]-\infty; -t[\cup]t; +\infty[$
- Il faut trouver t de sorte de contrôler l'erreur de première espèce
- t est le fractile d'ordre $1 - \alpha/2$ d'une loi normale

Exemple :
Si $\alpha = 0.05$, alors $t = 1.96$ et ainsi si $z_n = 2.4$ on rejette $\mathcal{H}_0$.

Comparaison de la moyenne à une norme : σ^2 connu (2)

Et si on n'avait pas émis l'hypothèse de loi normale sur notre variable d'intérêt X, est ce que cela aurait changé quelque chose?

Comparaison de la moyenne à une norme : σ^2 connu (2)

Et si on n'avait pas émis l'hypothèse de loi normale sur notre variable d'intérêt X, est ce que cela aurait changé quelque chose?

NON

Comparaison de la moyenne à une norme : σ^2 connu (2)

Et si on n'avait pas émis l'hypothèse de loi normale sur notre variable d'intérêt X, est ce que cela aurait changé quelque chose?

NON

En effet, grâce au théorème central limit, on aurait que $\bar{X}_n$ suit approximativement une loi $\mathcal{N}(\mu_0; \sigma = 0.05/\sqrt{n})$

Comparaison de la moyenne à une norme : σ^2 connu (2)

Méthode 2 :

- On calcule z_n

Comparaison de la moyenne à une norme : σ^2 connu (2)

Méthode 2 :

- On calcule z_n
- on calcule $P(|Z_n| \geq z_n)$: cette quantité est appelée la p-valeur

Comparaison de la moyenne à une norme : σ^2 connu (2)

Méthode 2 :

- On calcule z_n
- on calcule $P(|Z_n| \geq z_n)$: cette quantité est appelée la p-valeur
- si p-valeur $< \alpha$: on décide $\mathcal{H}_a$, sinon on ne rejette pas $\mathcal{H}_0$

Comparaison de la moyenne à une norme : σ^2 connu (2)

Méthode 2 :

- On calcule z_n
- on calcule $P(|Z_n| \geq z_n$: cette quantité est appelée la p-valeur
- si p-valeur $< \alpha)$: on décide $\mathcal{H}_a$, sinon on ne rejette pas $\mathcal{H}_0$

Exemple : Pour $z_n = 2.4$, on lit p-valeur $= 0.0164$. Puisque $p - valeur < \alpha$ on rejette $\mathcal{H}_0$

Comparaison de la moyenne à une norme : σ^2 inconnu

On a toujours $Z_n = \sqrt{n}\frac{\bar{X}_n - \mu_0}{\sigma} \sim \mathcal{N}(0; 1)$.

Comparaison de la moyenne à une norme : σ^2 inconnu

On a toujours $Z_n = \sqrt{n}\frac{\bar{X}_n - \mu_0}{\sigma} \sim \mathcal{N}(0; 1)$.

Mais Z_n dépend d'un paramètre inconnu

Comparaison de la moyenne à une norme : σ^2 inconnu

On a toujours $Z_n = \sqrt{n}\frac{\bar{X}_n - \mu_0}{\sigma} \sim \mathcal{N}(0;1)$.

Mais Z_n dépend d'un paramètre inconnu : σ

Comparaison de la moyenne à une norme : σ^2 inconnu

On a toujours $Z_n = \sqrt{n}\frac{\bar{X}_n - \mu_0}{\sigma} \sim \mathcal{N}(0; 1)$.

Mais Z_n dépend d'un paramètre inconnu : σ

Donc, on ne peut pas utiliser Z_n pour déterminer la zone de rejet

Comparaison de la moyenne à une norme : σ^2 inconnu

On a toujours $Z_n = \sqrt{n}\frac{\bar{X}_n - \mu_0}{\sigma} \sim \mathcal{N}(0;1)$.

Mais Z_n dépend d'un paramètre inconnu : σ

Donc, on ne peut pas utiliser Z_n pour déterminer la zone de rejet

On utilise la même astuce que dans la construction d'un intervalle de confiance à variance inconnue

Comparaison de la moyenne à une norme : σ^2 inconnu (2)

- on estime σ^2 par $\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

Comparaison de la moyenne à une norme : σ^2 inconnu (2)

- on estime σ^2 par $\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$
- si on a l'hypothèse de normalité sur X, alors on utilise le fait que $T_n = \sqrt{n}\frac{\bar{X}_n - \mu_0}{\hat{\sigma}_n} \sim \mathcal{T}(n-1)$

Comparaison de la moyenne à une norme : σ^2 inconnu (2)

- on estime σ^2 par $\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$
- si on a l'hypothèse de normalité sur X, alors on utilise le fait que $T_n = \sqrt{n}\frac{\bar{X}_n - \mu_0}{\hat{\sigma}_n} \sim \mathcal{T}(n-1)$
- sans hypothèse de normalité et avec n grand, alors on utilise le fait que $T_n = \sqrt{n}\frac{\bar{X}_n - \mu_0}{\hat{\sigma}_n} \sim \mathcal{N}(0;1)$

Comparaison de la moyenne à une norme : test unilatéral

Qu'est ce qui change si on procède non pas à un test bilatéral mais à un test unilatéral?

Comparaison de la moyenne à une norme : test unilatéral

Qu'est ce qui change si on procède non pas à un test bilatéral mais à un test unilatéral?

- test unilatéral supérieur ($\mathcal{H}_0 : \mu \leq \mu_0$)
  La zone de rejet est de la forme $z_n \in]t, +\infty[$ avec t le fractile d'ordre $1-\alpha$ d'une loi $\mathcal{N}(0;1)$ ou d'une loi $\mathcal{T}(n-1)$ selon le cadre

Comparaison de la moyenne à une norme : test unilatéral

Qu'est ce qui change si on procède non pas à un test bilatéral mais à un test unilatéral?

- test unilatéral supérieur ($\mathcal{H}_0 : \mu \leq \mu_0$)
  La zone de rejet est de la forme $z_n \in]t, +\infty[$ avec t le fractile d'ordre $1-\alpha$ d'une loi $\mathcal{N}(0;1)$ ou d'une loi $\mathcal{T}(n-1)$ selon le cadre

- test unilatéral inférieur ($\mathcal{H}_0 : \mu \geq \mu_0$)
  La zone de rejet est de la forme $z_n \in]-\infty, -t[$ avec t le fractile d'ordre $1-\alpha$ d'une loi $\mathcal{N}(0;1)$ ou d'une loi $\mathcal{T}(n-1)$ selon le cadre