IUT Génie Mécanique et Productique
Université Paul Sabatier Toulouse 3

Module M115 - Probabilités et statistiques
Partie 2/2 : Échantillonnage, estimation et tests d'hypothèse

1 Notions générales

1.1 Échantillonnage et intervalles de confiance

Considérons une population d'où l'on extrait un échantillon d'effectif n, dont les éléments sont notés x_i. La statistique descriptive associe à cet échantillone une valeur centrale, la moyenne empirique

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i$$

et une valeur de dispersion, la variance empirique

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2.$$

La loi de probabilité associée à cette population est inconnue, mais la moyenne empirique et la variance empirique fournissent des estimations de la moyenne et de la variance de la population. Lorsque la taille de l'échantillon s'accroît indéfiniment, les estimations tendent vers les valeurs exactes à estimer.

L'estimation des paramètres s'effectue à partir du seul échantillon, mais si on veut apprécier la qualité de cette estimation, il faut considérer la loi de probabilité attachée à la population. Elle est inconnue, et on suppose généralement qu'il s'agit de la loi normale. Si nous fixons un risque de se tromper en estimant par exemple la moyenne de la loi de probabilité de la population, par exemple 5%, les tables de la loi normale permettent de calculer la largeur d'un intervalle autour de la valeur vraie inconnue dans lequel il y a 95% de chances de trouver l'estimation. On peut alors considérer que cet intervalle représente les valeurs de la moyenne pour lesquelles la différence entre cette moyenne et l'observation n'est pas statistiquement significative au niveau 5%. En reportant cet intervalle autour de la valeur estimée, on dit qu'on a défini l'intervalle de confiance à 95% pour la moyenne.

1.2 Tests d'hypothèse et risques

Toute hypothèse concernant la loi de probabilité associée aux observations est une hypothèse statistique. On ne peut pas la vérifier, mais seulement la rejeter lorsque les observations paraissent en contradiction avec elle. Cependant, on ne pourra jamais affirmer avec certitude que l'hypothèse est fausse, mais seulement improbable. On se fixe donc a priori un risque α (probabilité de rejet de l'hypothèse qui serait réalisée malgré les apparences). La loi de probabilité de la grandeur considérée permet de déterminer une zone de probabilité $1-\alpha$, niveau de signification du test, dont le complément, de probabilité α, est appelé **région critique**. Si l'estimation tombe dans cette région critique, l'hypothèse doit être rejetée avec le risque α de se tromper. La **région d'acceptation** est la région complémentaire de la région critique. Elle correspond à l'intervalle dans lequel les différences observées entre les réalisations et la théorie sont attribuables aux fluctuations d'échantillonnage. La région

critique correspond donc aux intervalles dans lesquels les différences sont trop grandes pour être le fruit du hasard.

Un test d'hypothèse consiste en une succession d'étapes :
- énoncé de l'hypothèse principale H_0 et de l'hypothèse alternative H_1.
- calcul d'une variable de décision à partir de l'échantillon.
- calcul de la probabilité pour que H_0 soit vraie connaissant la variable calculée précédemment. Cette probabilité, appelée **risque de première espèce**, généralement notée α_0, correspond au risque de rejeter à tort H_0.
- conclution du test, en fonction d'un risque seuil arbitraire (on prend généralement 5%), en dessous duquel H_0 sera rejetée (c'est-à-dire que dans 5% des cas, l'expérimentateur se trompera quand il rejettera H_0).

La probabilité pour que H_0 soit fausse alors qu'on l'a acceptée est le **risque de deuxième espèce**, généralement notée β.

Le tableau résume les différents cas :

	H_0 vraie	H_1 vraie
H_0 décidée	$1-\alpha$	β
H_1 décidée	α	$1-\beta$

Ces deux erreurs sont antagonistes, plus α sera petit et plus β sera grand. En effet, le fait d'imposer α faible conduit à une règle de décision plus stricte qui aboutit le plus souvent à n'abandonner H_0 que dans des cas rarissimes et donc à conserver cette hypothèse quelques fois à tort.

On appelle **puissance d'un test** la quantité $1-\beta$.

2 Tests d'hypothèse

2.1 Test sur la moyenne à variance connue

On suppose que la loi de probabilité de la population est une loi normale de moyenne m inconnue et de variance σ^2 connue. On effectue généralement un test avec les hypothèses suivantes :

$$H_0 = \{m = m_0\} \text{ contre } H_1 = \{m < m_0\}.$$

Un bon estimateur de la moyenne m est la moyenne de l'échantillon, que l'on note

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Si chaque individu de l'échantillon suit une loi normale de paramètres m et σ^2, alors $\bar{X}$ suit une loi normale de paramètres m et $\dfrac{\sigma^2}{n}$, ce qui est équivalent à

$$\sqrt{n}\frac{\bar{X}-m}{\sigma} \text{ suit une loi } \mathcal{N}(0,1).$$

On va définir une règle de décision :
- Si $\bar{X} < \tilde{m}$, on rejette H_0 et on valide $m < m_0$.
- Si $\bar{X} > \tilde{m}$, on accepte H_0 et on valide $m = m_0$.

Pour déterminer le paramètre $\tilde{m}$ servant à décider quelle hypothèse doit être validée, on utilise la définition du risque de première espèce :

$$\alpha = P(\text{rejeter } H_0 \text{ quand } H_0 \text{ est vraie}) = P(\bar{X} < \tilde{m} \text{ quand } m = m_0).$$

Donc

$$\alpha = P\left(\sqrt{n}\frac{\bar{X} - m_0}{\sigma} < \sqrt{n}\frac{\tilde{m} - m_0}{\sigma}\right).$$

α étant donné, on en déduit la valeur de $\sqrt{n}\dfrac{\tilde{m} - m_0}{\sigma}$ dans une table de la loi normale centrée réduite. On en déduit ensuite la valeur de $\tilde{m}$. Il ne reste plus qu'à comparer la moyenne de l'échantillon avec cette valeur pour prendre une décision en fonction de la règle de décision.

2.2 Test sur la moyenne à variance inconnue

Lorsque la moyenne est inconnue, on utilise la variance observée s^2 et $\sqrt{n}\dfrac{\bar{X} - m}{s}$ suit alors une loi de Student T_{n-1} à $(n-1)$ degrés de liberté. On obtient alors

$$\alpha = P\left(T_{n-1} < \sqrt{n}\frac{\tilde{m} - m_0}{s}\right)$$

et pour α donné, on récupère $\tilde{m}$ grâce à une table de la loi de Student à $n-1$ degrés de liberté. La décision s'effectue de la même manière que dans le cas d'une variance connue.

2.3 Test sur la variance à moyenne connue

On considère les hypothèses $H_0 = \{\sigma^2 = \sigma_0^2\}$ et $H_1 = \{\sigma^2 < \sigma_0^2\}$. On obtient un bon estimateur de la variance σ^2 en calculant

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2.$$

Cet estimateur suit une loi du χ^2 à n paramètres (χ_n^2). On obtient alors, comme précédemment

$$\alpha = P\left(\chi_n^2 < \frac{(n-1)\tilde{\sigma}}{\sigma_0^2}\right)$$

où $\tilde{\sigma}$ est la paramètre permettant la décision : si $\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i - m)^2 < \tilde{\sigma}$, on rejette H_0 ; sinon on accepte H_0. La valeur de $\tilde{\sigma}$ se lit dans une table de la loi du χ^2 à n paramètres.

2.4 Test sur la variance à moyenne inconnue

Lorsque la moyenne est inconnue, on utilise le même procédé avec l'estimateur

$$\frac{(n-1)s^2}{\sigma^2}$$

(où s est la variance observée) qui suit une loi du χ^2 à $n-1$ paramètres. On a alors

$$\alpha = P\left(\chi_{n-1}^2 < \frac{(n-1)\tilde{\sigma^2}}{\sigma_0^2}\right)$$

ce qui permet de trouver $\tilde{\sigma^2}$. On rejette H_0 si $s^2 < \tilde{\sigma^2}$ et on accepte H_0 sinon.

2.5 Exemple

1. Un fabricant de peinture affirme que son nouveau modèle de peinture sèche en moyenne en $m = 20$ minutes. Une association de consommateurs voudrait prouver qu'en fait $m > 20$ mn. Un échantillon de 36 pots est testé et on mesure les temps de séchage. On suppose que le temps de séchage de la peinture suit une loi normale de moyenne m (inconnue) et de variance $\sigma^2 = 5,76$. On note $\bar{X}$ la moyenne observée. Le chargé détude de l'association met en place la règle suivante : si $\bar{X} > t$ mn, on rejette l'affirmation du fabriquant ; sinon on l'accepte. Déterminer t pour que le risque de conclure à tort que l'affirmation du fabriquant est fausse soit inférieur à 5%.
   Ce risque est le risque de première espèce, correspondant à la probabilité de choisir $H_1 = \{m > 20\}$ alors qu'en fait $H_0 = \{m = 20\}$ est vraie. Comme la variance est connue, on sait que $\sqrt{n}\dfrac{\bar{X} - m}{\sigma}$ suit une loi $\mathcal{N}(0,1)$. On a donc
   $$0.05 = P\left(\mathcal{N}(0,1) > \sqrt{36}\frac{t - 20}{2,4}\right)$$
   ce qui donne à l'aide d'une table de la loi normale
   $$\sqrt{36}\frac{t - 20}{2,4} = 1,65$$
   donc $t = 20,66$.
   On sait donc que si on laisse 40 minutes de marge par rapport aux affirmations du fabriquant, on a moins de 5% de chances de rejeter à tort l'affirmation du fabriquant.
2. En supposant que le fabriquant ment, et que la moyenne est en réalité $m = 21$, quelle est la probabilité que l'association ne s'en rende pas compte ?
   L'association va valider H_0, donc on cherche la probabilité d'avoir $\bar{X} < 20,66$ sachant que $m = 21$. C'est
   $$P\left(\mathcal{N}(0,1) < \sqrt{36}\frac{20,66 - 21}{2,4}\right) = P(\mathcal{N}(0,1) < -0.85) = 0,2$$
   donc il y a 20% de chances pour que l'association valide l'affirmation du fabriquant alors qu'en fait elle est fausse.

2.6 Tests bilatéraux

Les tests bilatéraux sont des tests du type $H_0 = \{m = m_0\}$ contre $H_1 = \{m \neq m_0\}$.

Exemple sur un test de moyenne à variance connue On sait que $\sqrt{n}\dfrac{\bar{X} - m}{\sigma}$ suit une loi $\mathcal{N}(0,1)$. On va chercher un intervalle centré autour de m_0, correspondant à la région d'acceptation de l'hypothèse H_0. Le risque de première espèce correspond alors à

$$\alpha = P(\text{rejeter } H_0 \text{ quand } H_0 \text{ est vraie}) = P(\bar{X} \notin [m_0 - t; m_0 + t] \text{ quand } m = m_0)$$
$$= P\left(\sqrt{n}\frac{\bar{X} - m_0}{\sigma} \notin \left[\frac{-\sqrt{n}t}{\sigma}; \frac{\sqrt{n}t}{\sigma}\right]\right).$$

α étant donné, il faut lire dans la table de la loi normale la valeur de $\dfrac{\sqrt{n}t}{\sigma}$, et donc de t.

On décidera alors de rejeter H_0 si $\bar{X} \notin [m_0 - t; m_0 + t]$, ou de l'accepter dans le cas contraire.