TICE Probabilités et statistiques 2 : Mémo de statistiques inférentielles

Présentation.

Il s'agit de donner des informations sur un paramètre théorique, qu'on notera m, à l'aide d'une mesure physique x, ou vice-versa. L'approche probabiliste de la mesure consiste à interpréter la valeur numérique x comme une valeur possible d'une variable aléatoire qu'on notera X, et dont la loi dépend de m. On supposera ici que l'espérance de X est égale à m.
Nous avons 2 façons de poser le problème :

- **Contrôler :** m est connu, et je veux savoir quelles valeurs raisonnables attendre de la mesure x : ce sont **les intervalles de fluctuations**. C'est le cas du contrôle de qualité, des tests de fiabilité ou plus généralement des tests d'hypothèses.
- **Estimer :** Je dispose de la mesure x, et je voudrais bien donner une estimation précise de la valeur théorique m. C'est dans ce cadre qu'apparaissent **les intervalles de confiance**.

1 Mesures

1.1 Fourchette de la mesure physique

Comme pour toutes les mesures physiques, le paramètre m ne peut être calculé exactement : sa valeur numérique n'est donc qu'une valeur approchée **à un terme d'erreur près** majoré par ϵ. Souvent, ce nombre est connu ou donné par la précision de l'appareil de mesure.
En réalité, on ne donne jamais une valeur du paramètre m, mais une fourchette de valeurs, de la forme $[x-\epsilon, x+\epsilon]$.
Remarquons que le rôle de m et x ici est symétrique, et que je peux donc les échanger : si le paramètre m est connu, je pourrais alors prévoir la fourchette du résultat x avant de l'avoir mesuré.

1.2 Cas des mesures statistiques

Dans le cadre de l'approche probabiliste, comme les valeurs de la variable aléatoire X fluctuent parfois dans un intervalle très grand, on rajoutera la notion de **degré de confiance** dans la fourchette proposée : il s'agit de la probabilité de notre fourchette soit correcte. Cela correspond au pourcentage théorique des cas où le résultat de la mesure est effectivement à moins de ϵ de sa valeur théorique. On dira dans ce cas que les valeurs des mesures trop loin de la valeur théorique sont des valeurs aberrantes.

2 Intervalles de fluctuations

2.1 Définition

On suppose que la valeur de m est connue. Dans ces conditions, si je réalise ou simule la mesure X, les valeurs que j'obtiendrais ne seront pas constantes. Elles vont fluctuer autour de m, avec une amplitude que je peux contrôler la plupart du temps : **un intervalle de fluctuation à 95%** est un intervalle qui contient théoriquement au moins 95% des résultats de la mesure X. En pratique, c'est un intervalle le plus petit possible de la forme $[a, b]$ où

$$P(X < a) \leq 2,5\% \quad \text{et} \quad P(X > b) \leq 2,5\%.$$

Lorsqu'on simule suffisamment de valeurs de la mesure X, environ au moins 95% des résultats de simulation se trouveront dans cet intervalle. Grâce à la loi des grands nombres, un autre moyen pour estimer un intervalle de fluctuation à 95% consiste à éliminer 5% des valeurs extrêmes des résultats de simulation.

2.2 Contrôle, fiabilité : principe du test d'hypothèse

Lorsqu'on veut contrôler la valeur du paramètre m on est amené à effectuer un test d'hypothèse : ce protocole est très courant dans des situations pratiques comme le contrôle de qualité, la fiabilité, mais aussi dans le domaine biologique et médical lorsqu'il s'agit par exemple d'étudier l'effet d'une substance sur un organisme (médicaments, ou au contraire substances toxiques).
L'hypothèse :
L'idée consiste à prendre pour valeur de m, la valeur "par défaut" (par exemple : pièce de monnaie non pipée, seuil de concentration maximal autorisé..). Cette hypothèse fixe la norme. Elle ne doit jamais être déterminée en fonction du résultat du test. Il s'agit d'un pré-requis. Cette règle est primordiale.

Le protocole de test :
Avec l'hypothèse, on définit un intervalle de fluctuation de X à 95% (qui s'obtient selon les cas en enlevant les valeurs extrêmes inférieures, ou supérieures, ou les 2) : dans le cadre de l'exemple d'une valeur maximale autorisée, l'intervalle de fluctuation n'excluera que les valeurs trop grandes.
Les conclusions :
La fiabilité du produit testé sera basée sur l'appartenance ou non de la mesure réalisée x à l'intervalle de fluctuation déterminé par les besoins. Il y a donc 2 cas possibles :
La mesure est dans l'intervalle de fluctuation : On dit alors que le produit est conforme au niveau 5% ; **mais attention :** il n'y a aucun contrôle du pourcentage de fraudeurs non détectés. Ce n'est pas parce que le produit passe les tests qu'il est effectivement dans les normes ("faux négatifs").
La mesure n'est pas dans l'intervalle de fluctuation : On concluera dans ce cas que le produit a 95% de chances de n'être pas conforme. Cela signifie que j'autorise de rejeter à tort 5% des produits respectant la norme (faux positifs).

3 Intervalles de confiance

3.1 Cadre

Dans ce paragraphe, l'objectif est de déterminer à l'aide du résultat x de la mesure X une fourchette de valeurs vraisemblables pour m. C'est le cas des sondages par exemple.
En général, les valeurs de X fluctuent autour de m, et un intervalle de fluctuation de X à 95% est de la forme $[m - \epsilon_m, m + \epsilon_m]$.

3.2 Principe-Définition

Dans notre cas, on ne dispose que de x, et on cherche un intervalle de valeurs possibles de m. **Un intervalle de confiance à 95% est l'ensemble de toutes les valeurs de m pour lesquelles x est une fluctuation "normale" de X (à 95%)** . Autrement dit, une valeur m de l'intervalle de confiance dès que x est dans l'intervalle de fluctuation à 95% pour cette valeur.
Pour calculer cet intervalle, on peut calculer sur machine les intervalles de fluctuation de X pour les différentes valeurs du paramètre m, et selectionner celles qui conviennent.

3.3 Un cas simple

Avec les notations ci-dessus, on remarque que x est dans l'intervalle de fluctuation à 95% si et seulement si $|x-m| \leq \epsilon_m$. En particulier, si $\epsilon_m = \epsilon$ est indépendant de m, alors on obtient que $m \in [x - \epsilon, x + \epsilon]$. Par conséquent :

si $[m - \epsilon, m + \epsilon]$ est un intervalle de fluctuation à 95% pour X,
alors $[x - \epsilon, x + \epsilon]$ est un intervalle de confiance à 95% pour m.

4 Quelques exemples

Nous explicitons dans les exemples qui suivent des intervalles de fluctuations et intervalles de confiance. Bien entendu, plus l'intervalle est petit, mieux c'est (la précision est meilleure).

4.1 Inégalité de Bienaymé-Chebichev

Théorème : *Si X est un variable aléatoire d'espérance m et de variance σ^2, alors on a*

$$P(|X - m| > \varepsilon) \leq \sigma^2/\varepsilon^2 \quad \textit{pour tout } \varepsilon > 0.$$

Pour déterminer un intervalle de fluctuation à 95%, il suffit que je choisisse ε tel que $\sigma^2/\varepsilon^2 = 5\%$. On obtient alors un intervalle de fluctuation de la forme $[m - 2\sigma\sqrt{5}, m + 2\sigma\sqrt{5}]$.

Lorsque σ ne dépend pas de m,
un intervalle de confiance à 95% pour m est de la forme $[x - 2\sigma\sqrt{5}, x + 2\sigma\sqrt{5}]$.

Ce n'est sûrement pas le plus précis, mais il marche.

4.2 Cas gaussien

Si X suit une loi gaussienne de paramètres (m, σ^2), alors $(X - m)/\sigma$ suit une loi gaussienne centrée réduite, et on peut calculer explicitement a pour que la quantité $P(|X - m|/\sigma > a) = 5\%$: on trouve a=1,96 (en valeur arrondie au centième), ce qui donne comme intervalle de fluctuation à 95% l'intervalle $[m - 1.96\sigma, m + 1.96\sigma]$, et comme intervalle de confiance correspondant l'intervalle $[x - 1.96\sigma, x + 1.96\sigma]$.

Intervalles de confiance des lois gaussiennes.
Intervalle de confiance pour m à 68% : $[x - \sigma, x + \sigma]$.
Intervalle de confiance pour m à 95% : $[x - 2\sigma, x + 2\sigma]$.
Intervalle de confiance pour m à à 99,8% : $[x - 3\sigma, x + 3\sigma]$.

4.3 Cas du sondage-intervalle asymptotique

4.3.1 Le modèle

Dans le cas d'un sondage portant sur 2 réponses de type "oui" ou "non", il s'agit d'estimer le pourcentage p de la population répondant "oui". Pour cela on dispose d'un échantillon de réponses, de taille n (de l'ordre de 1000 en général), et on note le nombre de oui obtenus. La variable X donnant intuitivement un résultat proche de p est la proportion de oui dans l'échantillon considéré. En notant X_i la variable de Bernoulli donnant 1 si la personne i répond "oui", on peut écrire $X = \sum_1^n X_i/n$, où on supposera que les X_i sont des variables aléatoires indépendantes de loi de Bernoulli de paramètre p. Dans ces conditions, nX suit une loi Binomiale de paramètre (n, p), et on peut donc déterminer par un calcul systématique l'intervalle de confiance de p selon les valeurs de x (le pourcentage observé).

4.3.2 Théorème de Moivre-Laplace

Le théorème de Moivre-Laplace dit que lorsque n est grand, la variable nX ressemble à une loi gaussienne de paramètres $(np, np(1-p))$. En pratique, on utilise cette approximation lorsque $n > 20$, et que nx et $n(1 - x)$ sont supérieurs à 5.

4.3.3 Intervalle de confiance asymptotique

Puisqu'en général, n est autour de 1000, dès que la proportion observée est comprise entre 2% et 98%, on obtient comme intervalle de fluctuations à 95% de X l'intervalle $[p - 2\sqrt{p(1-p)/n}, p + 2\sqrt{p(1-p)/n}]$. Majorons ensuite $p(1 - p)$ par 1/4, pour obtenir un intervalle de fluctuation dont la largeur ne dépend plus de p : le nouvel intervalle de fluctuation s'écrit $[p - 1/\sqrt{n}, p + 1/\sqrt{n}]$.

On obtient alors un intervalle de confiance asymptotique à 95% pour p de la forme $[x - 1/\sqrt{n}, x + 1/\sqrt{n}]$.