HAI718 Probabilités et statistiques

TD2 : Tests statistiques sur un paramètre d'intérêt

1 Introduction et problématique

Un industriel veut lancer sur le marché un nouveau produit issu de l'agriculture biologique que l'on note *Produit A*. Ce produit est bien entendu en concurrence avec d'autres produits.

Après une étude financière, les services comptables indiquent à cet industriel que pour le lancement d'un produit soit rentable, il faut qu'il soit vendu à plus de 300000 exemplaires par mois. La population ciblée par l'industriel est une population de taille $N = 2$ millions. Autrement dit, pour que le produit soit mis sur le marché, il faut que la proportion d'acheteurs potentiels soit supérieure à 15 %.

Exercice 1 *Dégagez le* **paramètre d'intérêt** *de l'étude et proposez-en une notation.*

1.1 La solution idéale

Exercice 2 *Quelle serait la solution idéale pour répondre à cette problématique ? Pourquoi cette solution paraît-elle peu envisageable ? Quelle serait une solution réalisable ?*

1.2 Une solution réalisable

La solution réalisable consiste à interroger un nombre moins important de consommateurs potentiels. Supposons que l'on puisse appliquer la procédure précédente sur une sous-population de taille $n = 100$ consommateurs potentiels.

Exercice 3 *Donnez une* **estimation** *du paramètre d'intérêt et une notation pour cette estimation ? En général, la valeur de cette estimation et la "vraie" valeur du paramètre d'intérêt sont différentes. Pourriez-vous en donnez les explications ?*

Comment conluriez-vous quant au problème posé par l'industriel ?

1.2.1 Formalisation mathématique du problème

- La population totale

Dans cette étude, la population totale correspond à la population des consommateurs potentiels. On la note $\Omega = \{\omega_1, \ldots, \omega_N\}$ où ω_i désigne le $i$ème individu de la population totale Ω.

Soit X la fonction qui à tout individu ω_i de la population totale associe la réponse de l'individu ω_i à la question : "achèteriez-vous le *Produit A* ?". En fait, on utilise le codage suivant : $X(\omega_i) = 1$ pour une réponse par **oui** et à $X(\omega_i) = 0$ pour une réponse par non.

Exercice 4 *Suivant ces notations, que vaut alors le paramètre d'intérêt ?*

- L'échantillon

La fonction X est associée à une *variable aléatoire* que l'on notera également X.

Exercice 5 *Que vaut $P(X=1)$? Quel est le nom de cette variable aléatoire et l'expression de sa densité ?*

On désigne par $\mathbf{e} = \{e_1, \ldots, e_n\}$ l'échantillon de taille n. À partir de celui-ci, on obtient alors le vecteur d'observations $\mathbf{x}^A = (x_1, \ldots, x_n)$ pour les variables $\mathbf{X}^A = (X_1, \ldots, X_n)$. Cela signifie simplement que $X(e_i) = x_i$ pour $i = 1, \ldots, n$.

Exercice 6 *Quel est l'estimateur du paramètre d'intérêt ? Quelle est son estimation ?*

1.3 Décision devant être prise

Exercice 7 *À l'aide des informations et des notations précédentes, proposez une règle de décision qui permettrait de donner une réponse à l'industriel. Argumentez et critiquez votre proposition. Qu'en pensez-vous ?*

1.4 Mise en évidence d'un phénomène aléatoire

Voici les résultats de trois études menées pour le lancement du produit dans trois pays différents. Dans chaque cas, l'étude est réalisée sur un échantillon de taille $n = 1000$ individus. Voici ce que l'on observe :

- **Pays 1** : le nombre de consommateurs potentiels est 123
- **Pays 2** : le nombre de consommateurs potentiels est 155
- **Pays 3** : le nombre de consommateurs potentiels est 176

Exercice 8 *La question à laquelle vous devez répondre est : Dans quel(s) pays l'industriel doit-il lancer le produit ?*

1.4.1 Décision prise par le statisticien

Le statisticien conseillerait de ne lancer son produit que pour le troisième pays, et de ne pas le lancer pour les pays 1 et 2.

Exercice 9 *Qu'en pensez-vous ? Quels pays auriez-vous choisi ?*

1.4.2 Les types d'erreur

Un agronome se procure des grains chez un fournisseur qui produit des sacs de deux qualités différentes : des sacs de bonne qualité qui contiennent 6% de grains "défectueux" (qui ne germent pas), et des sacs de mauvaise qualité contenant 10% de grains défectueux. On veut déterminer la qualité du sac reçu. On pose le test d'hypothèses suivant :

$$H_0 : p = 10\% \text{ contre } H_1 : p = 6\%.$$

L'effectif de l'échantillon pour réaliser le test est $n = 100$ grains. On note X_i une variable aléatoire de Bernoulli de paramètre p (notée $\mathcal{B}e(p)$) telle que :

$X_i = 1$ si le grain est défectueux et $P(X_i = 1) = p$
$X_i = 0$ si le grain n'est pas défectueux et $P(X_i = 0) = 1 - p$.

La loi de $X = \sum_{i=1}^{n} X_i$ est une loi Binomiale de paramètres n et p (notée $\mathcal{B}(n,p)$) avec :

$$P(X = k) = C_n^k p^k (1-p)^{(n-k)} \text{ et } C_n^k = \frac{n!}{k!(n-k)!}.$$

Pour $n > 30$, on peut approximer la loi de X/n par une loi normale et

$$X/n \sim \mathcal{N}\left(\mu = p, \sigma^2 = \frac{p(1-p)}{n}\right).$$

1.4.3 Les erreurs de première et seconde espèce

Avec un seuil de décision $s = 8\%$, on définit la règle de décision suivante :
si $\hat{p} \leq s$, on accepte H_1 et on abandonne H_0
si $\hat{p} > s$, on conserve H_0

Exercice 10 *Suivant cette règle de décision et en utilisant l'approximation de la loi X/n, quels sont les risques*

1. *de première espèce : $\alpha = P($accepter H_1 à tort$)$*
2. *de deuxième espèce : $\beta = P($conserver H_0 à tort$)$.*

Afin de répondre, on représentera graphiquement la distribution du test sous les deux hypothèses H_0 et H_1.

Admettons que l'on pose les hypothèses suivantes :

$$H_0 : p = 6\% \text{ contre } H_1 : p = 10\%.$$

Exercice 11 *Le résultat obtenu à la question précédente serait-il différent ? Expliquer pourquoi, éventuellement à l'aide de graphiques.*

1.5 Choix du risque α

Exercice 12 *On fixe $\alpha = 0.05$. Déterminer la valeur de s correspondante. Calculer la valeur du risque β. Qu'en concluez-vous ?*

1.6 Choix de α et β

Exercice 13 *On fixe $\alpha = 0.05$ et $\beta = 0.05$. Quelles sont les valeurs de s et de n correspondant à ces risques ?*

2 Le problème des naissances

≪ En 1660, le négociant anglais Graunt découvrait, d'après les registres des baptêmes anglicans, un rapport de 105 garçons pour 100 filles, rapport présentant deux étranges particularités : sa remarquable constance d'un pays à l'autre et d'une époque à l'autre et son accroissement après guerre ou les famines ! Double mystère longtemps attribué à un ≪ ordre divin ≫ mais qu'aucun chercheur n'a pu à ce jour éclaircir ≫ [1]

Voici ci-dessous des données recueillies par des étudiants du CUST (génie biologique) auprès de l'INSEE de Clermont-Ferrand.

	1969	1970	1971	1972
Allier : Garçons	2653	2735	2730	2716
Filles	2621	2521	2631	2640
Total	5274	5256	5361	5356
Cantal : Garçons	1290	1200	1257	1258
Filles	1268	1243	1246	1223
Total	2558	2443	2503	2481
Haute-Loire : Garçons	1504	1477	1469	1386
Filles	1437	1410	1443	1375
Total	2941	2887	2912	2761
Puy-de-Dôme : Garçons	4468	4567	4806	4864
Filles	4164	4263	4551	4664
Total	8632	8830	9357	9528
Auvergne : Garçons	9915	9979	10262	10224
Filles	9490	9437	9871	9902
Total	19405	19416	20133	20126

Au total 79080 naissances et parmi elles 40380 naissances de garçons.

Exercice 14 *Ces résultats confirment-ils ceux de Graunt ? À savoir :*

1. *Rejet de l'hypothèse qu'à la naissance il y a autant de filles que de garçons.*
2. *La proportion de garçons est de 105/205.*

3 Le problème ≪ Effet de tailles ≫

On jette une pièce de monnaie et on compte le nombre de fois où l'on obtient ≪ face ≫. On considère les trois cas suivants :

1. D'après Schwartz (1993) Le jeu de la Science et du Hasard. Ed. Flammarion.

- On jette 100 fois la pièce et on obtient 55 fois « Face ».
- On jette 1000 fois la pièce et on obtient 550 fois « Face ».
- On jette 10000 fois la pièce et on obtient 5500 fois « Face ».

Exercice 15 *Dans chacun de ces cas, peut-on accepter l'hypothèse selon laquelle la probabilité d'obtenir « Face » est 0,5 ?*