Fiche 12 - L3 MASS - Tests

Tran Viet Chi, `chi.tran@math.univ-lille1.fr`, bureau 316 (bâtiment M3).

Exercice 1 (Sondages)
Cet exercice se divise en plusieurs parties et se réalise sous `Scilab`.

On suppose qu'une proportion $p \in [0,1]$ de la population compte voter pour un candidat A tandis que les $1-p$ restants ont l'intention de voter pour un candidat B. On interroge $n = 1000$ personnes, choisies de façon indépendante dans la population, et on suppose qu'elles répondent honnêtement.
Partie A : estimation des intentions de vote

A chaque répondant $i \in [\![1, n]\!]$, on associe une variable aléatoire X_i qui vaut 1 s'il compte voter pour A et 0 s'il compte voter pour B. Ces variables aléatoires sont donc supposées *iid* de loi $\mathcal{B}(1,p)$.

1. Calculer l'estimateur du maximum de vraisemblance de p lorsqu'on a un échantillon $X_1, \dots, X_n$.

2. On considère la moyenne empirique $\bar{X}_n = \sum_{i=1}^n X_i/n$. Rappeler son espérance, sa variance et donner sa limite lorsque $n \to +\infty$. $\bar{X}_n$ est un estimateur de p.

3. Quelle est la loi de $\sum_{i=1}^n X_i$? Dessiner l'histogramme de $N = 3000$ simulations de variables *iid* de même loi que $\sum_{i=1}^n X_i$ pour $p = 0.5$. Dessiner en fonction de $p \in [0,1]$ la probabilité pour que $\bar{X}_n > 1/2$. Commenter.

4. Nous nous intéressons à la probabilité $\mathbb{P}(\bar{X}_n \notin]p-0.01, p+0.01[)$. Est-il possible de calculer explicitement cette probabilité ? Numériquement, dessiner cette probabilité en fonction de $p \in [0,1]$. Que vaut-elle pour $p = 1\%$, 50%, 75% ? Quelles sont les valeurs maximales et minimales ?

5. Pour $a > 0$, en utilisant l'inégalité de Bienaymé-Tchebychev, montrez que :

$$\mathbb{P}\left(\left|\bar{X}_n - p\right| \geq a\right) \leq \frac{1}{4a^2 n}$$

5.1. Quel est nombre d'individus n' à interroger pour que la probabilité que l'écart entre $\bar{X}_n$ et p soit supérieur à $a = 1\%$ soit inférieure à 5% ?
5.2. Donner un intervalle de confiance fonction de $\bar{X}_n$ et contenant p avec probabilité 0.95 lorsque l'on interroge $n = 1000$ personnes.

6. Lors d'un sondage réalisé par la TNS-Sofres le 24 avril 2007, 1000 personnes avaient été interrogées. Les intentions de vote étaient 51% Sarkozy contre 49% Royal. Commenter.

7. Pour les données de la question 6, faire sur l'ordinateur le test $H_0 \ : \ p = 0.5$ contre $H_1 \ : \ p > 0.5$.

Partie B : test de comparaison d'échantillons

On suppose qu'on dispose maintenant de 2 échantillons obtenus de façon indépendante et de tailles respectives n_1 et n_2 : $X_1, \dots X_{n_1}$ et $Y_1, \dots Y_{n_2}$. On suppose que les v.a. du premier échantillon sont des $\mathcal{B}(1,p_1)$ indépendantes et que les v.a. du second échantillon sont des $\mathcal{B}(1,p_2)$ indépendantes.

7. On souhaite tester $H_0 \ : \ p_1 = p_2$ contre $H_1 \ : \ p_1 \neq p_2$. On introduit la statistique

$$T_{n_1,n_2} = \frac{(\bar{X}_{n_1} - p_1) - (\bar{Y}_{n_2} - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}.$$

Quelle est la loi asymptotique de T_{n_1,n_2} lorsque n_1 et $n_2 \to +\infty$ avec $n_1/n_2 \to q > 0$?

8. A quelle statistique T'_{n_1,n_2} la statistique T_{n_1,n_2} est-elle égale sous H_0 ? Cette statistique est-elle calculable ?

9. Proposer une statistique T''_{n_1,n_2} calculable et donner sa loi sous H_0.

10. Quelle est le comportement de cette statistique sous H_1 ?

11. Dans un sondage réalisé entre le 18 avril 2012 sur 2552 personnes, l'Ifop demande "si dimanche prochain se déroulait le second tour de l'élection présidentielle, pour lequel des candidats suivants y aurait-il le plus de chances que vous votiez". 54% des personnes interrogées ont choisi Hollande contre 46% Sarkozy. Dans un sondage du CSA, le 17 avril 2012, à la question "Si le second tour de l'élection présidentielle de 2012 avait lieu dimanche prochain et que vous aviez le choix entre les deux candidats suivants, pour lequel y aurait-il le plus de chances que vous votiez ?", 58% des 886 personnes interrogées avaient choisi Hollande contre 42% Sarkozy.
Y a-t-il une différence significative entre ces deux résultats ?

Même question si l'on considère les sondages de la semaine précédente : le 16 avril 2012, l'Ifop publiait les scores de 55.5%-44.5% sur 1808 interrogés et le CSA trouvait les scores de 57%-43% sur 886 interrogés.

12. Donner la p-valeur des tests asymptotiques précédents.

13. On souhaiterait maintenant ne plus faire de tests asymptotiques, mais regarder ce qui se passe pour nos n_1 et n_2 donnés en question 11. Pour cela, comme on a des lois compliquées, on va faire reconstruire numériquement la loi de T''_{n_1,n_2} sous H_0 au lieu de la tabuler.

- Générer $N = 1000$ échantillons de tailles n_1 et n_2 sous H_0.
- Pour chaque simulation i, calculer $T''^{(i)}_{n_1,n_2}$. On a ainsi un ensemble $T''^{(1)}_{n_1,n_2}, \dots T''^{(N)}_{n_1,n_2}$ de réalisations de v.a. i.i.d. dont la loi est celle de T''_{n_1,n_2} sous H_0.
- Tracer l'histogramme de ces valeurs. Superposer une droite verticale indiquant la valeur T''_{n_1,n_2} observée sur les données de la question 11.
- Calculer numériquement la p-valeur et conclure.