Corrigé de l'exercice 6 de la feuille 4

Exercice 6 (de la feuille 4) Estimation d'une proportion

Sur un échantillon de 100 personnes, on constate que 60 sont fumeurs. Estimer la proportion des fumeurs dans la population. Déterminer un intervalle de confiance pour la proportion des fumeurs de niveau 95%.

Une correction possible . On choisit uniformément au hasard une personne dans la population totale. On note X la variable aléatoire égale à 1 si cette personne fume et 0 sinon. La variable X est une variable de Bernoulli, de paramètre p, où p est la proportion de fumeurs dans la population totale. Le but est de déterminer p.

On fait un échantillon de taille $n = 100$ personnes dans la population, et on note, comme précédemment, pour tout i entre 1 et n, $X_i = 1$ si la personne numéro i fume et 0 sinon.

On suppose que l'échantillon est constitué de tirages uniformes et indépendants avec remise, ou bien que la taille de la population totale est assez importante pour que l'on puisse faire comme si c'était le cas.

Finalement, X_1, X_2, ..., X_n sont i.i.d. de loi de Bernoulli de paramètre p inconnu.

Soit $\overline{X_n}$ la moyenne empirique des X_i, c'est-à-dire

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

D'après la loi des grands nombres, on a

$$\overline{X_n} \xrightarrow[n\to\infty]{p.s.} \mathbb{E}[X_1] = p.$$

Par conséquent, la moyenne empirique est un estimateur consistant du paramètre p. On peut aussi facilement montrer qu'il est sans biais. Dans l'exercice 3 (question 2), que malheureusement nous n'avons pas eu le temps de traiter, on montrer que c'est en fait l'estimateur du maximum de vraisemblance de p. Bref, tout porte à penser que c'est un bon estimateur de p.

D'après l'énoncé, une estimation de p est donc $\overline{X_n} = 60/100 = 0,6$.

Pour chercher un intervalle de confiance, on utilise le théorème central limite. On sait que la variance de chaque X_i est $p(1-p)$. Ainsi, on fait l'approximation que $\overline{X_n}$ suit la loi normale d'espérance p et de variance $p(1-p)/n$.

Le problème supplémentaire est que cette variance n'est pas connue non plus (elle dépend de p que l'on ne connaît pas). Mais d'après la loi des grands nombres, on sait aussi que

$$\overline{X_n}\left(1-\overline{X_n}\right) \xrightarrow[n\to\infty]{p.s.} p(1-p).$$

En utilisant ce fait et le théorème central limite, on a que

$$\frac{\sqrt{n}}{\sqrt{\overline{X_n}\left(1-\overline{X_n}\right)}}\left(\overline{X_n}-p\right) \xrightarrow[n\to\infty]{\text{(en loi)}} \mathcal{N}(0;1).$$

On fait l'approximation supplémentaire que le terme à gauche de la flèche a la loi normale centrée réduite pour faire les calculs.

On cherche un intervalle de confiance de niveau de confiance 95% donc on utilise la valeur $z_{0,975}$ qui est telle que, si Z suit la loi normale centrée réduite,

$$\mathbb{P}(Z \leq z_{0,975}) = 0,975.$$

D'après les tables (ou `norminv(0.975)` dans Octave), on a $z_{0,975} \approx 1,96$. Pour cette valeur, on a aussi, par symétrie,

$$\mathbb{P}(-1,96 \leq Z \leq 1,96) \approx 0,95.$$

Maintenant, on revient à notre estimation de p. Pour réduire la taille des formules, on note

$$\widehat{p} \coloneqq \overline{X_n}.$$

On a alors, avec les approximations faites précédemment,

$$\mathbb{P}\left(-1,96 \leq \frac{\sqrt{n}}{\widehat{p}\,(1-\widehat{p})}\,(\widehat{p}-p) \leq 1,96\right) \approx 0,95.$$

À l'intérieur des parenthèses, on fait les opérations suivantes :

1. on multiplie tous les termes de l'inégalité par $\frac{\widehat{p}(1-\widehat{p})}{\sqrt{n}}$ qui est positif ;
2. on soustrait $\widehat{p}$;
3. on multiplie par -1, qui est négatif, ce qui change le sens des inégalités.

Ceci fait, on obtient,

$$\mathbb{P}\left(\widehat{p}-1,96\frac{\widehat{p}(1-\widehat{p})}{\sqrt{n}} \leq p \leq \widehat{p}+1,96\frac{\widehat{p}(1-\widehat{p})}{\sqrt{n}}\right) \approx 0,95.$$

Donc notre intervalle de confiance au niveau de confiance 95% est

$$[\widehat{p}-1,96\frac{\widehat{p}\,(1-\widehat{p})}{\sqrt{n}}\,;\widehat{p}+1,96\frac{\widehat{p}\,(1-\widehat{p})}{\sqrt{n}}].$$

C'est l'intervalle de confiance, classique, utilisé pour estimer une proportion. Souvent, les statisticiens disent que c'est valable si : $n \geq 30$, $n\widehat{p} \geq 5$ et $n(1-\widehat{p}) \geq 5$.

Dans notre cas, ces trois conditions sont vérifiées. On fait l'application numérique. On trouve

$$1,96\frac{\widehat{p}(1-\widehat{p})}{\sqrt{n}} = \frac{1,96 \times 0,6 \times 0,4}{\sqrt{100}} = 0,04704.$$

L'intervalle est donc

$$[0,553; 0,647]\,.$$

On peut dire que la *vraie* proportion p est dans cette intervalle avec un risque d'erreur de 5%.