

Statistiques : Enoncé du TP 3 Sondages

Dans une population de grande taille, une élection oppose un candidat A à une candidate B. On désigne par p le pourcentage de personnes qui voteront pour A et par $1 - p$ celui des personnes qui voteront pour B (on suppose pour simplifier qu'il n'y a ni abstention ni votes blancs ou nuls). Par un sondage effectué auprès de n personnes, on *estime* la valeur p par la proportion f de ces n personnes qui votera pour A. Pour cela, on prélève au hasard dans la population un *échantillon* de n individus (n est la *taille* de l'échantillon), et on calcule la fréquence (dite fréquence empirique ou *fréquence estimée*) du caractère "voter pour A" dans cet échantillon (nombre de personnes interrogées favorables à A divisé par n).

Dans ce TP, nous allons étudier comment cette fréquence estimée varie en fonction de l'échantillon (*fluctuation d'échantillonnage*), puis étudier la distribution de ses valeurs, voir de quelle façon l'on construit l'*intervalle de confiance* de cette estimation et enfin étudier comment ces fluctuations diminuent lorsque la taille de l'échantillon augmente.

1. Fluctuations d'échantillonnage : Pour simuler un sondage auprès d'un échantillon de $n = 400$ personnes, dans une population où $p = 0.4$ est la proportion des personnes favorables au candidat A et $1 - p = 0.6$ celle des personnes favorables à la candidate B, on simule une suite de 400 v.a. de Bernoulli $\mathcal{B}(1, 0.4)$. On obtient ainsi un vecteur de 400 composantes, égales à 0 ou 1, le 1 signifiant que la personne sondée votera pour le candidat A et le 0 qu'elle votera pour B.

1. L'instruction suivante permet de simuler un tel sondage auprès de 400 personnes :

```
sondage=int(rand(1,n)+p)
```

Avant de saisir cette instruction, taper `n=400;p=0.4` ; puis examiner le sens et la syntaxe de la fonction `int` à l'aide de l'aide en ligne. A la suite de cette instruction, faites le calcul de la fréquence empirique f_e du caractère "Voter pour A" dans l'échantillon simulé.

Répéter la simulation et le calcul de f_e plusieurs fois pour observer les fluctuations de f_e lorsqu'on change d'échantillon.

2. La fonction `grand` permet de simuler plusieurs lois dont la loi Binomiale (et donc en particulier la loi de Bernoulli). En vous aidant de l'aide en ligne, trouver une autre façon de simuler un tel sondage, en utilisant `grand` et non plus `rand`.

3. Pour étudier les fluctuations d'échantillonnage, on simule, en une fois, 100 sondages, et pour chacun on calcule la fréquence empirique f_e par la suite d'instructions suivantes :

```
sondages=int(rand(100,n)+p);fe=sum(sondages,'c')/n;
```

On obtient ainsi une distribution de 100 valeurs de f_e que l'on va étudier. Tracer un histogramme en 15 classes de cette distribution.

2. Position et étendue de la distribution des valeurs de f_e :

1. Calculer les deux paramètres de position, moyenne et médiane.
2. Calculer les trois paramètres d'étendue, plus petite moins plus grande valeurs), écart-type et intervalle interquartile (pour ce dernier on pourra utiliser les instructions

```
q=quart(fe);q(3)-q(1)
```

3. L'instruction suivante

```
entreab=fe(a<=fe & fe<b);nbentreab=length(entreab)
```

permet de calculer le nombre de composantes du vecteur f_e qui appartient à l'intervalle $[a, b[$. Calculer ce nombre pour $[a, b[= [Min(fe), Max(fe)[$, puis pour $[a, b[= [q(1), q(3)[$.

4. La quantité (notée σ) $\sigma = \sqrt{p(1-p)/n}$ est l'écart type théorique de f_e . La calculer et la comparer à l'écart type empirique obtenu.

3. Intervalle de confiance :

1. Chaque fréquence estimée f_e est le centre d'un intervalle de confiance qui, en principe, doit contenir la fréquence exacte. Lorsque $\alpha = 5\%$, cet intervalle est égal à $[f_{einf}, f_{esup}]$, où $f_{einf} = f_e - 1,96\sigma$ et $f_{esup} = f_e + 1,96\sigma$. Définir les trois vecteurs `f_einf`, `f_theo` et `f_esup`, où `f_theo=p*ones(100,1)`.
2. La commande
`i=1 :100 ;plot2d(i,f_einf(i));plot2d(i,f_theo(i));plot2d(i,f_esup(i));`
trace les extrémités des intervalles de confiance et la fonction constante égale à p . Sauvegarder cette figure.
3. Calculer le nombre d'intervalles de confiance qui ne contiennent pas la valeur exacte (ou théorique) de p .
4. Reprendre les 3 questions précédentes pour $\alpha = 10\%$.

4. Influence de la taille de l'échantillon : Pour vérifier que les intervalles de confiance sont bien plus *resserrés* lorsque la taille de l'échantillon est plus importante, on va reprendre l'étude précédente avec $n = 1000$.

1. Effectuer un sondage dans ce cas puis calculer son intervalle de confiance au seuil de $\alpha = 5\%$, puis au seuil de $\alpha = 10\%$.
2. Effectuer 100 sondages et représenter l'histogramme des valeurs obtenues. Sauvegardez votre figure.
3. Calculer ses moyennes théoriques et empiriques.
4. Calculer ses écarts type théoriques et empiriques.
5. Pour $\alpha = 5\%$, puis pour $\alpha = 10\%$, calculer le nombre de "bons" sondages (sondages pour lesquels l'intervalle de confiance contient la valeur exacte).
6. Refaire ce dernier calcul pour $n = 10000$.