Corrigé Statistiques inférentielle par par Pierre Veuillez

1 Intervalle de confiance.

Exercice

Déterminer une valeur approchée de la loi de la moyenne empirique :
$E\left(\overline{X_n}\right) = E\left(X\right)$, $V\left(\overline{X_n}\right) = \frac{1}{n}V\left(X\right)$ donc $\overline{X_n} \underset{\sim}{\hookrightarrow} \mathcal{N}\left(E\left(X\right), \frac{1}{n}V\left(X\right)\right)$

2 Exercices

2.1 Variance

Soit X ayant une espérance m et une variance v, sa **variance empirique** est $W_n = \frac{1}{n}\sum X_i^2 - \overline{X_n}^2$ avec $\overline{X_n}$ la moyenne empirique de X et $\frac{1}{n}\sum X_i^2$ la moyenne empirique de X^2.

1. Soit Y ayant une espérance et une variance. Calculer $E\left(Y^2\right)$ en fonction $E\left(Y\right)$ et $V\left(Y\right)$
2. Calculer $E\left(\overline{X_n}\right)$ et $V\left(\overline{X_n}\right)$ et en déduire $E\left(\overline{X_n}^2\right)$
3. Montrer enfin que $E\left(W_n\right) = \frac{n-1}{n}V\left(X\right)$ et en déduire un estimateur sans biais de la variance.

Solution

1. $V\left(Y\right) = E\left(Y^2\right) - E\left(X\right)^2$ donc $E\left(Y^2\right) = V\left(Y\right) + E\left(Y\right)^2$
2. $E\left(\overline{X_n}\right) = m$ et $V\left(\overline{X_n}\right) = \frac{1}{n}v$ donc $E\left(\overline{X_n}^2\right) = \frac{1}{n}v + m^2$
3. $E\left(W_n\right) = \frac{1}{n}\sum E\left(X_i^2\right) = \frac{1}{n}n\left(v + m^2\right) - \left(\frac{1}{n}v + m^2\right) = \left(1 - \frac{1}{n}\right)v = \frac{n-1}{n}v$
   D'où $E\left(\frac{n-1}{n}W_n\right) = v$ et $\frac{n-1}{n}W_n$ variance empirique sans biais est un estimateur sans biais de la variance.

2.2 Question confidentielle.

Certains sujets abordés dans les enquêtes d'opinion sont parfois assez intimes, et on court le risque que les personnes interrogées se refusent à répondre franchement à l'enquêteur, faussant ainsi le résultat.
On peut alors avoir recours à une astuce consistant à inverser aléatoirement les réponses .
Considérons une question confidentielle pour laquelle on veut estimer la probabilité p de réponses positives.
L'enquêteur demande à chaque personne interrogée de lancer un dé.

- Si le dé tombe sur , la personne doit donner sa réponse sans mentir,
- sinon elle doit donner l'opinion contraire à la sienne.

Si l'enquêteur ignore le résultat du dé, il ne pourra pas savoir si la réponse est franche ou non, et on peut espérer que la personne sondée acceptera de jouer le jeu.
Généralisons légèrement la situation en tirant pour chaque personne une variable de Bernoulli de paramètre α.

- Si le résultat de cette variable est 1, la réponse est franche,
- sinon, elle est inversée.

Soit n le nombre de personnes interrogées.
L'enquêteur ne recueille que la fréquence empirique F_n des "oui".

1. Montrer que la probabilité de "oui" à l'issue de la procédure est $q = \alpha\, p + (1-\alpha)\,(1-p)$
2. Montrer que F_n, la fréquence observée par l'enquêteur, est un estimateur sans biais de q et de risque quadratique tendant vers 0 quand n tend vers $+\infty$
3. Pour $\alpha \neq 1/2$ exprimer p en en fonction de q.
4. En déduire que $T_n = \frac{F_n - 1 + \alpha}{2\alpha - 1}$ est un estimateur sans biais de p dont le risque quadratique tend vers 0 quand n tend vers $+\infty$.
5. Pour n fixé, quelle valeur attribuer à α pour que le risque quadratique soit minimum ? Est-ce acceptable ?

   Pour quelle valeur de α ce risque est-il maximum ?

   Quel sera le risque quadratique avec le dé ($\alpha = 1/6$)

2.3 Loi uniforme

Soit X de loi $\mathcal{U}[0,a]$ et $(X_1, \dots X_n)$ une n-échantillon de variables. Etimation de a :
X a une espérance de $a/2$. Soit $\overline{X_n}$ la moyenne empirique.

1. Soit $T_n = 2\overline{X_n}$. Montrer que T_n est sans biais et déterminer son risque quadratique
2. Soit $T'_n = \max(X_1, \dots, X_n)$

   Déterminer la fonction de répartition de X puis celle de T'_n

   En déduire sa densité puis son biais et son risque quadratique.
3. Soit $T''_n = \frac{n+1}{n} T'_n$ déterminer son biais et son risque quadratique.
4. Quel est le meilleur estimateur de a pour de grandes valeurs de n ?

solution:

1. $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$ donc $E\left(\overline{X_n}\right) = \frac{1}{n}\sum_{i=1}^n E(X_i) = \frac{a}{2}$ d'où $E(T_n) = 2\frac{a}{2} = a$ et T_n est sans biais.
   $V\left(\overline{X_n}\right) = \frac{1}{n^2}\sum_{i=1}^n V(X_i)$ car les (X_i) sont indépendantes.
   $E(X_i^2) = \int_0^a \frac{1}{a} t^2\, dt = \frac{1}{a}\left[t^3/3\right]_0^a = \frac{a^2}{3}$ donc $V(X_i) = \frac{a^2}{3} - \frac{a^2}{4} = \frac{a^2}{12}$ d'où $V\left(\overline{X_n}\right) = \frac{n\,a^2}{12n^2}$.
   La variance de $T_n = 2\overline{X_n}$ est alors $V(T_n) = 4V\left(\overline{X_n}\right) = \frac{a^2}{3n}$ et donc son risque quadratique est $\frac{a^2}{3n} + 0^2 = \frac{a^2}{3n}$
2. La fonction de répartition F de Xest : $F(x) = \int_{-\infty}^x f(t)\, dt = \begin{cases} 0 & \text{si } x < 0 \\ \int_0^x \frac{1}{a} dt = \frac{x}{a} & \text{si } x \in [0,a] \\ 1 & \text{si } x > a \end{cases}$
   $(T'_n \leq t) = (\max(X_1, \dots, X_n) \leq t) = \bigcap_{i=1}^n (X_i \leq t)$ et en notant F la fonction de répartition de X, et G celle de T'_n on a alors

$G(t) = F(t)^n$.

F est continue sur $\mathbb{R}$ et C^1 sauf en 0 et a donc G également et T'_n est à densité de densité :

$$g(t) = G'(t) = n\, f(t)\, F^{n-1}(t) = \begin{cases} 0 & \text{si } x \notin [0,a] \\ \frac{n}{a}\left(\frac{x}{a}\right)^{n-1} & \text{si } x \in [0,a] \end{cases}$$

L'espérance (qui existe) de T'_n est alors $\int_0^a t\, g(t)\, dt = \int_0^a \frac{n}{a^n} t^n dt = \left[\frac{n}{n+1}\frac{1}{a^n}t^{n+1}\right]_0^a = \frac{n}{n+1}a$

Donc T'_n a pour biais $\left(\frac{n}{n+1} - 1\right)a = -\frac{a}{n}$ (biaisé mais son biais tend vers 0 quand $n \to +\infty$)

L'espérance (qui existe) de ${T'_n}^2$ est $\int_0^a t^2\, g(t)\, dt = \int_0^a \frac{n}{a^n} t^{n+1} dt = \left[\frac{n}{n+2}\frac{1}{a^n}t^{n+2}\right]_0^a = \frac{n}{n+2}a^2$

Donc la variance de T'_n est

$$V(T'_n) = E\left({T'_n}^2\right) - E(T'_n)^2 = \frac{n}{n+2}a^2 - \left(\frac{n}{n+1}\right)^2 a^2 = \frac{n}{(n+1)^2(n+2)}a^2$$

et son risque quadratique est $r' = V(T'_n) + b^2 = \frac{n}{(n+1)^2(n+2)}a^2 + \frac{1}{n^2}a^2 = \left(\frac{n}{(n+1)^2(n+2)} + \frac{1}{n^2}\right)a^2 \sim \frac{2}{n^2}a^2$

3. Alors $T''_n = \frac{n+1}{n}T'_n$ a pour espérance $\frac{n+1}{n}E(T'_n) = a$ donc T''_n est sans biais.

   Sa variance est $V(T''_n) = \left(\frac{n+1}{n}\right)^2 V(T'_n) = \frac{1}{n(n+2)}a^2$ et a pour risque quadratique $r'' = \frac{1}{n(n+2)}a^2 \sim \frac{1}{n^2}a^2$ ce qui est (pour n grand) deux fois mieux que T'_n.

4. Donc pour de grandes valeurs de n, T''_n est le meilleur estimateur de a.

2.4 Intervalle de confiance pour le paramètre d'une variable de Bernouilli.

Lors d'un sondage sur 100 personnes interrogée, 60 pensent voter pour A
On modélise le choix par un échantillon $(X_1, \ldots, X_{100})$ de variable indépendantes de même loi de Bernouilli de paramètre p.
On cherche à déterminer un intervalle de confiance pour p au niveau de confiance 99% (1% de risque)

1. Déterminer l'espérance et la variance de la fréquence empirique $F = \frac{1}{100}\sum_{i=1}^{100} X_i$?

2. On note F^* la fréquence empirique centrée réduite.

   Par quelle loi peut on approcher celle de F^*? On suppose désormais que F^* suit $\mathcal{N}(0,1)$

3. Déterminer t tel que $\mathrm{P}(-t \leq F^* \leq t) \geq 0,99$ et en déduire que $\mathrm{P}\left(F - t\frac{\sqrt{p(1-p)}}{10} \leq p \leq F + t\frac{\sqrt{p(1-p)}}{10}\right) \geq 0,99$

4. Montrer que pour tout $p \in [0,1]$, $p(1-p) \leq \frac{1}{4}$ et en déduire que $[F - t/20\,;\, F + t/20]$ est un intervalle de confiance de p au niveau de confiance 99%

Solution

1. On a $E(F_{100}) = E\left(\frac{1}{100}\sum_{i=1}^{100} X_i\right) = \frac{1}{100}\sum_{i=0}^{100} E(X_i) = \frac{1}{100}100p = p$
   Donc F_n est un estimateur sans biais de p

2. Somme de variables indépendantes de même loi $\mathcal{B}(1,p)$: $V(X_i) = p(1-p) \neq 0$ et $E(X_i) = p$
   Donc avec $F = \frac{1}{100}\sum_{i=1}^{100} X_i$, F^* peut être approchée par une loi Normale centrée réduite.
   $V(F) = \frac{1}{100^2}\sum_{i=1}^{100} V(X_i)$ car les $(X_i)_i$ sont indépendantes. Donc $V(F) = \frac{1}{100}p(1-p)$ et
   $F^* = \frac{F-p}{\sqrt{\frac{p(1-p)}{100}}} = \frac{10}{\sqrt{p(1-p)}}(F-p)$ la fréquence empirique centrée réduite suit approximativement une loi $\mathcal{N}(0,1)$

3. Comme $-t \leq t$: $\mathrm{P}(-t \leq F^* \leq t) = \Phi(t) - \Phi(-t) = \Phi(t) - (1-\Phi(t)) = 2\Phi(t) - 1$
   On résout : $2\Phi(t) - 1 = 0,99 \iff \Phi(t) \geq 0,995$ et on lit sur la table de la lo Normale pour $t = 2,58$

   N.B. première transformation à connaître :

$$\begin{aligned}(-t \leq F^* \leq t) &= \left(-t \leq \frac{10}{\sqrt{p(1-p)}}(F-p) \leq t\right)\\ &= \left(-t\frac{\sqrt{p(1-p)}}{10} \leq F-p \leq t\frac{\sqrt{p(1-p)}}{10}\right)\\ &= \left(F - t\frac{\sqrt{p(1-p)}}{10} \leq p \leq F + t\frac{\sqrt{p(1-p)}}{10}\right)\end{aligned}$$

   Donc $\mathrm{P}\left(F_n - t\frac{\sqrt{p(1-p)}}{10} \leq p \leq F_n + t\frac{\sqrt{p(1-p)}}{10}\right) \geq 0,99$

4. On étudie les variations de $f(p) = p(1-p)$.
   f est dérivable sur $\mathbb{R}$ et $f'(p) = 1-p-p = 1-2p$

p	0		1/2		
$f'(p) = 1-2p$		+	0	−	affine
$f(p)$		↗	1/4	↘	

   et $p(1-p) \leq \frac{1}{4}$

   On a alors $\sqrt{p(1-p)} \leq \frac{1}{2}$ donc

   N.B. seconde transformation à connaître :
   $\left(F_n - t\frac{\sqrt{p(1-p)}}{10} \leq p \leq F_n + t\frac{\sqrt{p(1-p)}}{10}\right) \subset \left(F_n - t\frac{1}{20} \leq p \leq F_n + t\frac{1}{20}\right)$ et $\mathrm{P}(F_n - t/20 \leq p \leq F_n + t/20) \geq$
   $\mathrm{P}\left(F_n - t\frac{\sqrt{p(1-p)}}{20} \leq p \leq F_n + t\frac{\sqrt{p(1-p)}}{20}\right) \geq 0,99$

   Donc $[F_n - t/20\,;\ F_n + t/20]$ est un intervalle de confiance de p au niveau de confiance 99% soit avec l'échantillon de données : $\hat{p} = 0,6$

   $t/20 \simeq 0,13$, l'intervalle de confiance au niveau 99% est $[0,47\,;\,0,73]$ ... ce qui ne renseigne pas beaucoup sur les chances de remporter l'élection..

   Avec un échantillon de taille 10000, on trouvera l'intervalle $[F_n - t/200\,,\ F_n + t/200]$ soit une largeur d'intervalle proche de 5% pour un niveau de confiance de 99%.

Avec un niveau de confiance de 95%, on a $t = 1,96$ et pour $n = 1000$ on a $t\frac{\sqrt{p(1-p)}}{\sqrt{1000}} \leq 0,0302$, c'est la classique des sondages : pour un échantillon de 1000 personne, le résultat est donné avec un intervalle de confiance de 3% (ce que ne disent pas les sondeurs, c'est que cela n'est sûr qu'à 95% : il y a 5% de chance que la valeur réelle soit hors de cet intervalle de

2.5 Intervalle de confiance par Bienaymé-Tchebichev

Soit $a \in \left[0; 2\sqrt{3}\right]$, $X \hookrightarrow \mathcal{U}_{[0,a]}$ et $(X_1 \dots X_n)$ un n-echantillon de variables de même loi que X et indépendantes.
On cherche un intervalle de confiance de $\frac{a}{2}$ au niveau de confiance 99% (niveau de risque 1%).
On note $\overline{X_n}$ la moyenne empirique

1. Rappeler la moyenne m de X et montrer que $V(X) = \frac{a^2}{12}$. En déduire la moyenne et l'espérance de $\overline{X_n}$.
2. En déduire que $\mathrm{P}\left(\left|\overline{X_n} - \frac{a}{2}\right| > 0,1\right) \leq \frac{100}{n}$
3. Déterminer enfin n pour que $\left[\overline{X_n} - 0,1\,;\, \overline{X_n} + 0,1\right]$ soit un intervalle de confiance de $\frac{a}{2}$ au niveau de confiance 99%
4. Ecrire un programme PASCAL qui
   - choisit un nombre a au hasard dans $\left[0; 2\sqrt{3}\right]$
   - effectue 10000 tirages dans $[0, a]$
   - calcule et affiche la moyenne des résultats obtenus.

   Le programme a affiché 0,534.
   - Pensez vous que $\frac{a}{2} = 0,534$?
   - Pensez vous que $\frac{a}{2} > 0,7$?
   - Pensez vous que $\frac{a}{2} \in [0,43\,;\, 0,64]$?
5. Par quelle loi peut-on approcher celle de $\overline{X_{1000}}$?
6. Déterminer t pour que $\mathrm{P}\left(-t \leq \frac{\sqrt{12}}{a}100\left(\overline{X_n} - \frac{a}{2}\right) < t\right) \geq 0,99$ et en déduire un autre intervalle de confiance de $\frac{a}{2}$ au niveau α

Solution

Soit $a \in \left[0; 2\sqrt{3}\right]$, $X \hookrightarrow \mathcal{U}_{[0,a]}$ et $(X_1 \dots X_n)$ un $n$échantillon de variables de même loi que X et indépendantes.
On cherche un intervalle de confiance de $\frac{a}{2}$ au niveau de confiance 99% (niveau de risque 1%).
On note $\overline{X_n}$ la moyenne empirique

1. On a $E(X) = \frac{a}{2}$

   Et comme la densité de X est nulle hors de $[0, a]$ et vaut $\frac{1}{a}$ sur $[0, a]$ on a $E(X^2) = \int_0^a \frac{t^2}{a}dt = \left[\frac{t^3}{3a}\right]_0^a = \frac{a^2}{3}$ et donc X a une variance qui est $V(X) = \frac{a^2}{3} - \left(\frac{a}{2}\right)^2 = \frac{a^2}{12}$

   Donc $E\left(\overline{X_n}\right) = E\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n E(X_i) = \frac{n}{n}E(X) = \frac{a}{2}$

Et $V\left(\overline{X_n}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V\left(X_i\right)$ car les X_i sont indépendants $\cdots = \frac{1}{n^2} nV\left(X\right) = \frac{a^2}{12n}$

Rappeler la moyenne m de X et montrer que $V\left(X\right) = \frac{a^2}{12}$. En déduire la moyenne et l'espérance de $\overline{X_n}$.

2. D'après l'inégalité de Bienaymé-Tchebichev on a alors $\mathrm{P}\left(\left|\overline{X_n} - \frac{a}{2}\right| > 0,1\right) \leq \frac{V\left(\overline{X_n}\right)}{0,1^2} = 100\frac{a^2}{12n}$ et comme $0 \leq a \leq 2\sqrt{3}$ alors $a^2 \leq 12$ et donc $\mathrm{P}\left(\left|\overline{X_n} - \frac{a}{2}\right| > 0,1\right) \leq \frac{100}{n}$ et $\mathrm{P}\left(\left|\overline{X_n} - \frac{a}{2}\right| \leq 0,1\right) \geq 1 - \frac{100}{n}$

3. Comme l'événement $\left(\left|\overline{X_n} - \frac{a}{2}\right| \leq 0,1\right)$ s'écrit $\left(-0,1 \leq \overline{X_n} - \frac{a}{2} \leq 0,1\right)$ ou encore : $\left(\overline{X_n} - 0,1 \leq \frac{a}{2} \leq \overline{X_n} + 0,1\right)$

   Donc pour $n = 10000$ on a $\mathrm{P}\left(\overline{X_n} - 0,1 \leq \frac{a}{2} \leq \overline{X_n} + 0,1\right) \geq 1 - 0,01$ et $\left[\overline{X_n} - 0,1\,;\, \overline{X_n} + 0,1\right]$ est un intervalle de confiance de $\frac{a}{2}$ au niveau de confiance 99%

4. Ecrire un programme PASCAL qui

```
Program estim;
var a,x,s:real;k:integer;
begin
     randomize;a:=random(2*(3));s:=0;{initialisation}
     for k:=1 to 10000 do
     begin
          x:=random(a);
          s:=s+x;
     end;
     writeln('la moyenne est :',s/10000);
end.
```

   Le programme a affiché 0,534.

   - Chaque valeur a une probabilité nulle d'avoir été choisie ! donc $\frac{a}{2} \neq 0,534$?
   - La probabilité que $\frac{a}{2}$ soit dans l'intervalle $[0,534 - 0,1\ ;\ 0,534 + 0,1]$ est supérieure à 99%. Donc la probabilité qu'il soit $> 0,7$ est de moins de 1%. Je ne pense donc pas que $a/2 > 0,7$
   - La probabilité de $\frac{a}{2} \in [0,43\,;\,0,64]$ est supérieure à 99%. Je pense donc que oui. (et j'ai moins de 1% de chances de me tromper ...)

5. La loi $\sum_{i=1}^{n} X_i$ somme de variables indépendantes de même loi qui a pour espérance $n\frac{a}{2}$, et pour variance $n\frac{a^2}{12}$.

   DOnc centreée réduite, elle peut être approchée par une loi $\mathcal{N}\left(0,1\right)$ et $\overline{X_n}^* = \frac{\overline{X_n} - a/2}{\sqrt{a^2/12n}}$ par $\mathcal{N}\left(0,1\right)$

6. Et pour $n = 10000$: $\mathrm{P}\left(-t \leq \frac{\sqrt{12}}{a} 100\left(\overline{X_n} - \frac{a}{2}\right) < t\right) \simeq \Phi\left(t\right) - \Phi\left(-t\right) = 2\Phi\left(t\right) - 1$

   On résout $2\Phi\left(t\right) - 1 \geq 0,99 \Longleftrightarrow \Phi\left(t\right) \geq 0,995$ ce qui est vérifié pour $t = 2,58 \leq 2,6$

   On a $\left(-t \leq \frac{\sqrt{12}}{a} 100\left(\overline{X_n} - \frac{a}{2}\right) < t\right) = \left(\overline{X_n} - t\frac{a}{100\sqrt{12}} \leq \frac{a}{2} < \overline{X_n} + t\frac{a}{100\sqrt{12}}\right)$ avec $\frac{a}{100\sqrt{12}} \leq \frac{1}{100}$

donc $\left[\overline{X_n} - 0,026\,;\, \overline{X_n} + 0,026\right]$ est un intervalle de confiance de $\frac{a}{2}$ au niveau de confiance 99% (soit une précision quatre fois meilleure qu'avec la formule de Bienaymé-Tchebichev)

2.6 Comptage par capture et recapture

On cherche à évaluer le nombre N de poissons dans un étang.
Pour cela, on prélève dans l'étang m poissons que l'on bague avant les remettre dans l'étang.
On propose deux méthodes différentes d'estimation de N.

Méthode 1
Soit $n \in \mathbb{N}^*$, $n \geq m$.
On prélève des poissons dans l'étang, au hasard et avec remise.
On note X_n la variable aléatoire égale au nombre de poissons qu'il a été nécessaire de pêcher pour obtenir n poissons marqués.
Pour tout $i \in [2, n]$, on pose $D_i = X_i - X_{i-1}$. On pose $D_1 = X_1$ et on suppose que les D_i sont des variables indépendantes.

1. a) Pour tout $i \in [2, n]$, quelle est la signification de D_i ?
   b) Déterminer, pour $i \in [2, n]$, la loi de D_i, son espérance et sa variance.
   En déduire l'espérance et la variance de X_n.
   c) On pose $A_n = \frac{m}{n} X_n$. Montrer que A_n est un estimateur sans biais de N et déterminer son risque quadratique.
2. a) Pour n assez grand, par quelle loi peut-on approcher la loi de la variable aléatoire $\overline{X_n} = \frac{X_n}{n}$ (on utilisera le théorème de la limite centrée)?
   b) On a marqué 200 poissons puis effectué 450 prélèvements pour obtenir 50 poissons marqués.
   On pose $\sigma = \sigma(A_n)$. On a pu prouver par ailleurs que $\sigma \leq 100$.
   Déterminer en fonction de σ, un intervalle de confiance pour N au seuil 0.9 (On donne $\Phi(1,64) \simeq 0,95$).

Méthode 2
On prélève successivement et avec remise n poissons. Soit Y_n le nombre de poissons marqués parmi eux.

1. Montrer que $\frac{1}{nm} Y_n$ est un estimateur sans biais de $\frac{1}{N}$.
2. Pour quelle raison évidente ne peut-on pas prendre $\frac{nm}{Y_n}$ comme estimateur de N ?
   On pose alors $B_n = \frac{m(n+1)}{Y_n+1}$
   a) Calculer l'espérance de B_n.
   b) Est-il un estimateur sans biais de N ?

Solution

Méthode 1

1. a) D_i est la différence du nombre de pèche nécessaire pour obtenir $i-1$ et i poissons marqués.
   C'est le nombre de pèche pour obtenir un poisson marqué de plus.

b) Donc D_i est le **nombre de** pèches pour obtenir un poisson marqué de plus dans une **suite** de pèche (on peut supposer que la pèche se continue indéfiniment) **indépendantes** (avec remise, en supposant que les poissons sont bêtes et ne se souviennent pas qu'il ne faut pas mordre à l'hameçon) ayant toutes une probabilité $\frac{m}{N}$ de donner un poisson marqué.
Donc $D_i \hookrightarrow \mathcal{G}\left(\frac{m}{N}\right)$ et $E(D_i) = \frac{N}{m}$ et $V(D_i) = \dfrac{1-\frac{m}{N}}{\left(\frac{m}{N}\right)^2} = \dfrac{N(N-m)}{m^2}$
Comme $D_1 + D_2 + \cdots + D_n = X_n$ on a alors $E(X_n) = n\frac{N}{m}$ et comme les $(D_i)_i$ sont indépendants, $V(X_n) = n\frac{N(N-m)}{m^2}$

c) On pose $A_n = \frac{m}{n}X_n$.
On a alors $E(A_n) = \frac{m}{n}E(X_n) = N$ donc A_n est un estimateur sans biais de N.
Sa variance est $V(A_n) = V\left(\frac{m}{n}X_n\right) = \frac{m^2}{n^2}V(X_n) = \frac{N(N-m)}{n}$
Donc son risque quadratique est : $\text{biais}^2 + V(A_n) = \frac{N(N-m)}{n}$

2. a) Pour n assez grand, X_n étant une somme de variables indépendantes et de même loi, X_n^* peut être approchée par une loi normale centrée réduite.

b) A_n suit alors également une loi normale de paramètres $E(A_n) = N$ et et $V(A_n) = \sigma^2$ et $\frac{A_n - N}{\sigma}$ suit une loi normale centrée réduite.
Donc $\mathrm{P}\left(-t \leq \frac{A_n-N}{\sigma} \leq t\right) = \Phi(t) - \Phi(-t) = \Phi(t) - [1 - \Phi(t)] = 2\Phi(t) - 1$
Et

$$\begin{aligned} \mathrm{P}\left(-t \leq \frac{A_n - N}{\sigma} \leq t\right) \geq 0,9 &\iff 2\Phi(t) - 1 \geq 0,9 \\ &\iff \Phi(t) \geq 0,95 \simeq \Phi(1,64) \\ &\iff t \geq 1,64 \end{aligned}$$

car Φ est croissante sur $\mathbb{R}$
Comme $\sigma \leq 100$ alors

$$\left(-t \leq \frac{A_n - N}{\sigma} \leq t\right) = (A_n - t\sigma \leq N \leq A_n + t\sigma) \subset (A_n - t100 \leq N \leq A_n + t100)$$

Et avec $t = 1,64$: $\mathrm{P}(A_n - t100 \leq N \leq A_n + t100) \geq \mathrm{P}\left(-t \leq \frac{A_n-N}{\sigma} \leq t\right) \geq 0,9$
Donc $[A_n - 164\,,\, A_n + 164]$ est un intervalle de confiance de N au niveau de confiance $0,9$
Avec ici : $m = 200$; $n = 50$ et $X_{50} = 450$
Donc $A_{50} = \frac{200}{50}X_{50} = 1800$ (Estimation ponctuelle de N)
et on est sûr à 90% que le nombre de poissons dans l'étang est compris dans l'intervalle $[1636\,,\, 1964]$

Méthode 2
On prélève successivement et avec remise n poissons. Soit Y_n le nombre de poissons marqués parmi eux.

1. Le nombre Y_n de poissons marqués suit une loi binomial de paramètres $\left(n, \frac{m}{N}\right)$.
Donc son espérance est $E(Y_n) = n\frac{m}{N}$ et $E\left(\frac{1}{nm}Y_n\right) = \frac{1}{N}$
Donc $\frac{1}{nm}Y_n$ est un estimateur sans biais de $\frac{1}{N}$.
On a $V(Y_n) = n\frac{m}{N}\left(1 - \frac{m}{N}\right) = \frac{nm(N-m)}{N^2}$ donc $V\left(\frac{1}{nm}Y_n\right) = \left(\frac{1}{nm}\right)^2 V(Y_n) = \frac{(N-m)}{n\,m\,N^2}$
Donc le risque quadratique de $\frac{1}{nm}Y_n$ comme estimateur de $\frac{1}{N}$ est $\frac{(N-m)}{n\,m\,N^2}$

2. Comme Y_n peut être nul avec une probabilité non nulle, $\frac{nm}{Y_n}$ aurait une probabilité non nulle de ne pas être défini.
On pose alors $B_n = \frac{m(n+1)}{Y_n+1}$

a) On utilise le théorème de transfert : les valeurs de Y_n sont $[[0, n]]$

$$\begin{aligned} E(B_n) &= \sum_{k=0}^{n} \frac{m(n+1)}{k+1} \mathrm{P}(Y_n = k) \\ &= \sum_{k=0}^{n} \frac{m(n+1)}{k+1} \binom{n}{k} p^k q^{n-k} \end{aligned}$$

il faut développer le coefficient du binôme pour simplifier l'expression.
en notant $p = \frac{m}{N}$ et $q = 1 - \frac{m}{N}$

$$\begin{aligned} E(B_n) &= \sum_{k=0}^{n} \frac{m(n+1)}{k+1} \frac{n!}{k!\,(n-k)!} p^k q^{n-k} \\ &= \sum_{k=0}^{n} m \frac{(n+1)!}{(k+1)!\,(n-k)!} p^k q^{n-k} \end{aligned}$$

On y reconnaît $\binom{n+1}{k+1}$ et on réindexe $h = k + 1$pour faire réapparaitre la formule du binôme... pour la puissance $n + 1$

$$\begin{aligned} E(B_n) &= \sum_{k=0}^{n} m \binom{n+1}{k+1} p^k q^{n-k} \\ &= \sum_{k=1}^{n+1} m \binom{n+1}{h} p^{h-1} q^{n+1-h} \\ &= \frac{m}{p} \left(\sum_{k=0}^{n+1} \binom{n+1}{h} p^h q^{n+1-h} - q^{n+1} \right) \\ &= \frac{m}{p} \left((p+q)^{n+1} - q^{n+1} \right) \\ &= \frac{m}{p} \left(1 - q^{n+1} \right) \\ &= N \left(1 - q^{n+1} \right) \end{aligned}$$

b) Donc B est biaisé, mais quand n tend vers $+\infty$ (quand on augmente le nombre de repêche) le biais tend vers 0 : il est asymptotiquement sans biais.