Dénombrement et échantillonnage

Laurent Delsol (*MAPMO, Université d'Orléans.*)

Centre Galois,
21/06/2013, Orléans.

(laurent.delsol@univ-orleans.fr)

Motivations générales

Dans de nombreuses situations on cherche à dénombrer le nombre d'individus, d'animaux, de végétaux ou d'objets se trouvant dans une zone géographique donnée.

- décompte exact des individus souvent très difficile, voire impossible (grande étendue, population est trop importante).
- méthodes d'échantillonnage $\rightarrow$ approximation de la taille de la population en n'observant que :
  - certains individus
  - certaines zones géographiques.

L'objectif de cet atelier :

- se familiariser avec ces méthodes au travers de quelques exemples.
- présenter les résultats mathématiques sur lesquels elles reposent.

Plan de l'atelier

Observations répétées de phénomènes aléatoires

Chaque participant lance 6 fois 5 dés et note les résultats.
Calculer le score moyen et le nombre de 5 ou 6 observés.

- Comparer les résultats obtenus par chacun d'entre vous. D'où proviennent ces différences ?
- Lorsqu'on lance un dé :
  - quelle est la probabilité que l'on tombe sur une certaine face ?
  - quelle est la probabilité de tomber sur 5 ou 6 ?
  - quel est le score moyen auquel on peut s'attendre ?
- Quelle est la proportion globale de 5 ou 6 observés ?
- Quelle est la moyenne globale des scores obtenus ?
- Que remarquez-vous ?

Observations répétées de phénomènes aléatoires

Chaque participant lance 6 fois 5 dés et note les résultats.
Calculer le score moyen et le nombre de 5 ou 6 observés.

- Comparer les résultats obtenus par chacun d'entre vous. D'où proviennent ces différences ?
- Lorsqu'on lance un dé :
  - quelle est la probabilité que l'on tombe sur une certaine face ?
  - quelle est la probabilité de tomber sur 5 ou 6 ?
  - quel est le score moyen auquel on peut s'attendre ?
- Quelle est la proportion globale de 5 ou 6 observés ?
- Quelle est la moyenne globale des scores obtenus ?
- Que remarquez-vous ?

Observations répétées de phénomènes aléatoires

Chaque participant lance 6 fois 5 dés et note les résultats.
Calculer le score moyen et le nombre de 5 ou 6 observés.

- Comparer les résultats obtenus par chacun d'entre vous. D'où proviennent ces différences ?
- Lorsqu'on lance un dé :
  - quelle est la probabilité que l'on tombe sur une certaine face ?
  - quelle est la probabilité de tomber sur 5 ou 6 ?
  - quel est le score moyen auquel on peut s'attendre ?
- Quelle est la proportion globale de 5 ou 6 observés ?
- Quelle est la moyenne globale des scores obtenus ?
- Que remarquez-vous ?

Observations répétées de phénomènes aléatoires

Chaque participant lance 6 fois 5 dés et note les résultats.
Calculer le score moyen et le nombre de 5 ou 6 observés.

- Comparer les résultats obtenus par chacun d'entre vous. D'où proviennent ces différences ?
- Lorsqu'on lance un dé :
  - quelle est la probabilité que l'on tombe sur une certaine face ?
  - quelle est la probabilité de tomber sur 5 ou 6 ?
  - quel est le score moyen auquel on peut s'attendre ?
- Quelle est la proportion globale de 5 ou 6 observés ?
- Quelle est la moyenne globale des scores obtenus ?
- Que remarquez-vous ?

Observations répétées de phénomènes aléatoires

Chaque participant lance 6 fois 5 dés et note les résultats.
Calculer le score moyen et le nombre de 5 ou 6 observés.

- Comparer les résultats obtenus par chacun d'entre vous. D'où proviennent ces différences ?
- Lorsqu'on lance un dé :
  - quelle est la probabilité que l'on tombe sur une certaine face ?
  - quelle est la probabilité de tomber sur 5 ou 6 ?
  - quel est le score moyen auquel on peut s'attendre ?
- Quelle est la proportion globale de 5 ou 6 observés ?
- Quelle est la moyenne globale des scores obtenus ?
- Que remarquez-vous ?

Observations répétées de phénomènes aléatoires

Chaque participant lance 6 fois 5 dés et note les résultats.
Calculer le score moyen et le nombre de 5 ou 6 observés.

- Comparer les résultats obtenus par chacun d'entre vous. D'où proviennent ces différences ?
- Lorsqu'on lance un dé :
  - quelle est la probabilité que l'on tombe sur une certaine face ?
  - quelle est la probabilité de tomber sur 5 ou 6 ?
  - quel est le score moyen auquel on peut s'attendre ?
- Quelle est la proportion globale de 5 ou 6 observés ?
- Quelle est la moyenne globale des scores obtenus ?
- Que remarquez-vous ?

Observations répétées de phénomènes aléatoires

Chaque participant lance 6 fois 5 dés et note les résultats.
Calculer le score moyen et le nombre de 5 ou 6 observés.

- Comparer les résultats obtenus par chacun d'entre vous. D'où proviennent ces différences ?
- Lorsqu'on lance un dé :
  - quelle est la probabilité que l'on tombe sur une certaine face ?
  - quelle est la probabilité de tomber sur 5 ou 6 ?
  - quel est le score moyen auquel on peut s'attendre ?
- Quelle est la proportion globale de 5 ou 6 observés ?
- Quelle est la moyenne globale des scores obtenus ?
- Que remarquez-vous ?

Observations répétées de phénomènes aléatoires

Chaque participant lance 6 fois 5 dés et note les résultats.
Calculer le score moyen et le nombre de 5 ou 6 observés.

- Comparer les résultats obtenus par chacun d'entre vous. D'où proviennent ces différences ?
- Lorsqu'on lance un dé :
  - quelle est la probabilité que l'on tombe sur une certaine face ?
  - quelle est la probabilité de tomber sur 5 ou 6 ?
  - quel est le score moyen auquel on peut s'attendre ?
- Quelle est la proportion globale de 5 ou 6 observés ?
- Quelle est la moyenne globale des scores obtenus ?
- Que remarquez-vous ?

Loi des grands nombres

Ce phénomène, appelé loi des grands nombre est à la base de nombreuses méthodes statistiques.

Théorème

Soient n observations aléatoires et indépendantes $(X_1, \ldots, X_n)$ *d'un phénomène X de valeur moyenne m.*
Sous des hypothèses générales on montre que

$$\frac{(X_1 + \cdots + X_n)}{n} \underset{n \to +\infty}{\to} m. \qquad (\mathbb{P}, p.s.)$$

En d'autres termes, si n est assez grand :

- la valeur moyenne de X peut être approchée par la moyenne de n observations indépendantes.
- la probabilité d'un événement peut être approchée par la proportion de fois où il a été observé au cours de n observations indépendantes.

Loi des grands nombres

Ce phénomène, appelé loi des grands nombre est à la base de nombreuses méthodes statistiques.

Théorème

Soient n observations aléatoires et indépendantes $(X_1, \ldots, X_n)$ *d'un phénomène X de valeur moyenne m.*
Sous des hypothèses générales on montre que

$$\frac{(X_1 + \cdots + X_n)}{n} \underset{n \to +\infty}{\to} m. \qquad (\mathbb{P}, p.s.)$$

En d'autres termes, si n est assez grand :

- la valeur moyenne de X peut être approchée par la moyenne de n observations indépendantes.
- la probabilité d'un événement peut être approchée par la proportion de fois où il a été observé au cours de n observations indépendantes.

Loi des grands nombres

Ce phénomène, appelé loi des grands nombre est à la base de nombreuses méthodes statistiques.

Théorème

Soient n observations aléatoires et indépendantes $(X_1, \ldots, X_n)$ d'un phénomène X de valeur moyenne m.
Sous des hypothèses générales on montre que

$$\frac{(X_1 + \cdots + X_n)}{n} \underset{n \to +\infty}{\to} m. \qquad (\mathbb{P}, p.s.)$$

En d'autres termes, si n est assez grand :

- la valeur moyenne de X peut être approchée par la moyenne de n observations indépendantes.
- la probabilité d'un événement peut être approchée par la proportion de fois où il a été observé au cours de n observations indépendantes.

Limite centrale et intervalles de confiance

Notations :

moyenne : $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$

variance : $\sigma^2_{n-1} = \frac{1}{n-1}((X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2)$.

renormalisation : $T = (\overline{X} - m)/\sqrt{\frac{\sigma^2_{n-1}}{n}}$.

Construites à partir d'observations de phénomènes aléatoires, nos approximations présentent une certaine variabilité.

- Celle-ci diminue lorsque le nombre d'observations n croît.
- Pour n assez grand, les valeurs de T se répartissent à peu près suivant une loi $\mathcal{N}(0,1)$ [th. de la limite centrale].
- $P(\overline{X} - \sqrt{\frac{\sigma^2_{n-1}}{n}} \times 1.96 \leq m \leq \overline{X} + \sqrt{\frac{\sigma^2_{n-1}}{n}} \times 1.96) \approx 0.95$.

Limite centrale et intervalles de confiance

Notations :

moyenne : $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$

variance : $\sigma^2_{n-1} = \frac{1}{n-1}((X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2)$.

renormalisation : $T = (\overline{X} - m)/\sqrt{\frac{\sigma^2_{n-1}}{n}}$.

Construites à partir d'observations de phénomènes aléatoires, nos approximations présentent une certaine variabilité.

- Celle-ci diminue lorsque le nombre d'observations n croît.
- Pour n assez grand, les valeurs de T se répartissent à peu près suivant une loi $\mathcal{N}(0,1)$ [th. de la limite centrale].
- $P(\overline{X} - \sqrt{\frac{\sigma^2_{n-1}}{n}} \times 1.96 \leq m \leq \overline{X} + \sqrt{\frac{\sigma^2_{n-1}}{n}} \times 1.96) \approx 0.95$.

Limite centrale et intervalles de confiance

Notations :

moyenne : $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$

variance : $\sigma^2_{n-1} = \frac{1}{n-1}((X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2)$.

renormalisation : $T = (\overline{X} - m)/\sqrt{\frac{\sigma^2_{n-1}}{n}}$.

Construites à partir d'observations de phénomènes aléatoires, nos approximations présentent une certaine variabilité.

- Celle-ci diminue lorsque le nombre d'observations n croît.
- Pour n assez grand, les valeurs de T se répartissent à peu près suivant une loi $\mathcal{N}(0,1)$ [th. de la limite centrale].
- $P(\overline{X} - \sqrt{\frac{\sigma^2_{n-1}}{n}} \times 1.96 \leq m \leq \overline{X} + \sqrt{\frac{\sigma^2_{n-1}}{n}} \times 1.96) \approx 0.95$.

Limite centrale et intervalles de confiance

Notations :

moyenne : $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$

variance : $\sigma^2_{n-1} = \frac{1}{n-1}((X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2)$.

renormalisation : $T = (\overline{X} - m)/\sqrt{\frac{\sigma^2_{n-1}}{n}}$.

Construites à partir d'observations de phénomènes aléatoires, nos approximations présentent une certaine variabilité.

- Celle-ci diminue lorsque le nombre d'observations n croît.
- Pour n assez grand, les valeurs de T se répartissent à peu près suivant une loi $\mathcal{N}(0,1)$ [th. de la limite centrale].
- $P(\overline{X} - \sqrt{\frac{\sigma^2_{n-1}}{n}} \times 1.96 \leq m \leq \overline{X} + \sqrt{\frac{\sigma^2_{n-1}}{n}} \times 1.96) \approx 0.95$.

Limite centrale et intervalles de confiance

Notations :

moyenne : $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$

variance : $\sigma_{n-1}^2 = \frac{1}{n-1}((X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2)$.

renormalisation : $T = (\overline{X} - m)/\sqrt{\frac{\sigma_{n-1}^2}{n}}$.

Construites à partir d'observations de phénomènes aléatoires, nos approximations présentent une certaine variabilité.

- Celle-ci diminue lorsque le nombre d'observations n croît.
- Pour n assez grand, les valeurs de T se répartissent à peu près suivant une loi $\mathcal{N}(0,1)$ [th. de la limite centrale].
- $P(\overline{X} - \sqrt{\frac{\sigma_{n-1}^2}{n}} \times 1.96 \leq m \leq \overline{X} + \sqrt{\frac{\sigma_{n-1}^2}{n}} \times 1.96) \approx 0.95$.

Limite centrale et intervalles de confiance

Notations :

moyenne : $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$

variance : $\sigma^2_{n-1} = \frac{1}{n-1}((X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2)$.

renormalisation : $T = (\overline{X} - m)/\sqrt{\frac{\sigma^2_{n-1}}{n}}$.

Construites à partir d'observations de phénomènes aléatoires, nos approximations présentent une certaine variabilité.

- Celle-ci diminue lorsque le nombre d'observations n croît.
- Pour n assez grand, les valeurs de T se répartissent à peu près suivant une loi $\mathcal{N}(0,1)$ [th. de la limite centrale].
- $P(\overline{X} - \sqrt{\frac{\sigma^2_{n-1}}{n}} \times 1.96 \leq m \leq \overline{X} + \sqrt{\frac{\sigma^2_{n-1}}{n}} \times 1.96) \approx 0.95$.

Limite centrale et intervalles de confiance

Notations :

moyenne : $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$

variance : $\sigma_{n-1}^2 = \frac{1}{n-1}((X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2)$.

renormalisation : $T = (\overline{X} - m)/\sqrt{\frac{\sigma_{n-1}^2}{n}}$.

Construites à partir d'observations de phénomènes aléatoires, nos approximations présentent une certaine variabilité.

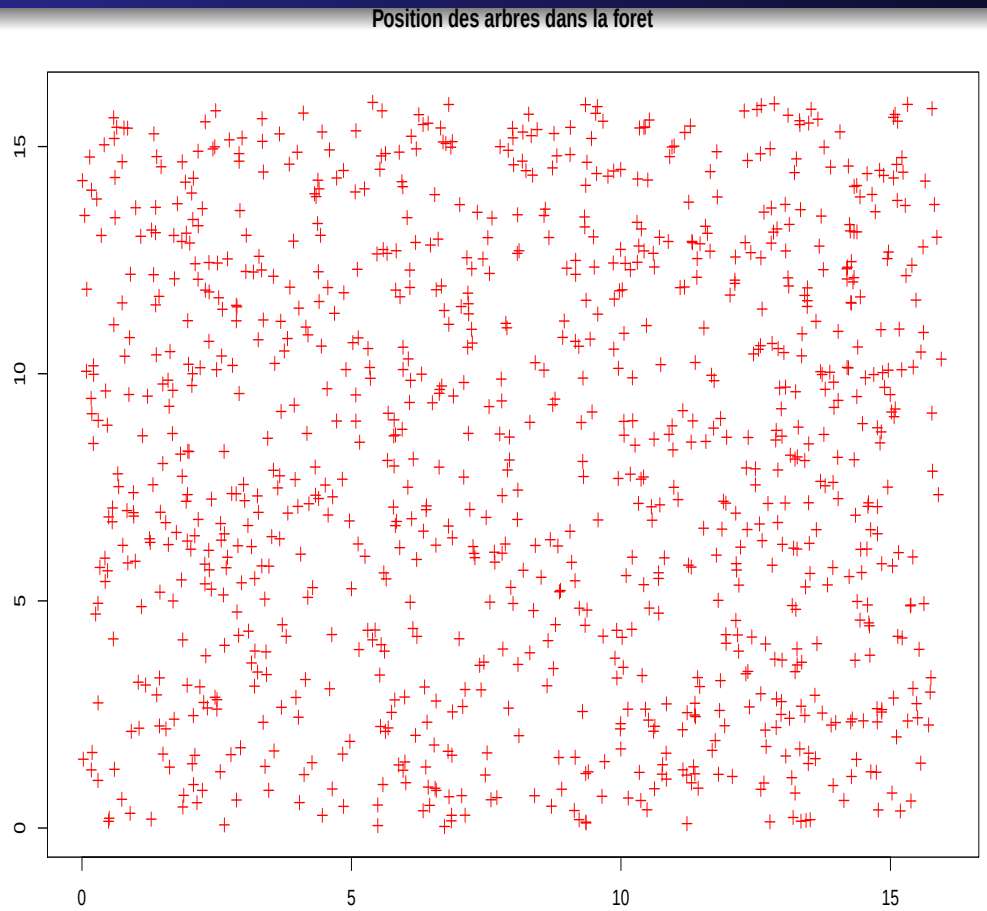
- Celle-ci diminue lorsque le nombre d'observations n croît.
- Pour n assez grand, les valeurs de T se répartissent à peu près suivant une loi $\mathcal{N}(0,1)$ [th. de la limite centrale].
- $P(\overline{X} - \sqrt{\frac{\sigma_{n-1}^2}{n}} \times 1.96 \leq m \leq \overline{X} + \sqrt{\frac{\sigma_{n-1}^2}{n}} \times 1.96) \approx 0.95$.

Echantillonnage aléatoire

Problème concret

On cherche à évaluer le nombre N d'érables se trouvant dans une forêt s'étendant sur une grande superficie.

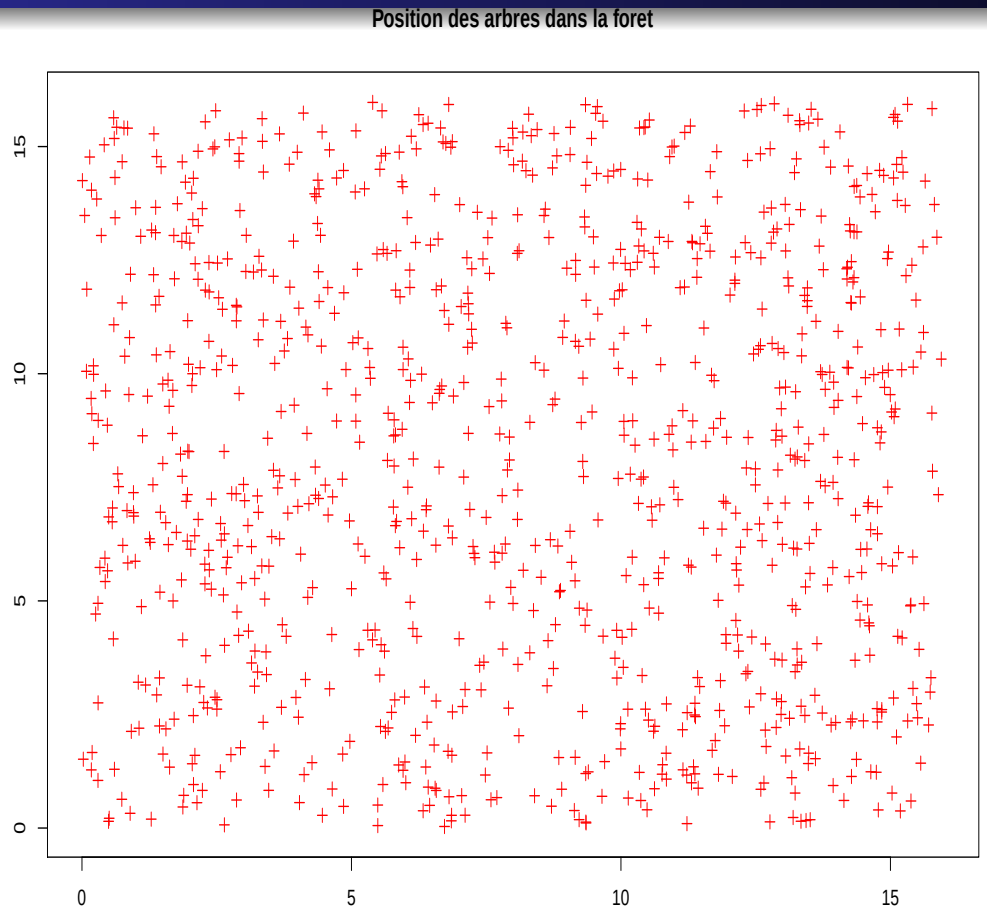
Comment faire ?



Pensez-vous que cela soit envisagable de dénombrer

- tous les érables de cette forêt ?
- les érables se trouvant dans différentes zones d'assez petite taille ?

Comment faire ?



Pensez-vous que cela soit envisagable de dénombrer

- tous les érables de cette forêt ?
- les érables se trouvant dans différentes zones d'assez petite taille ?

Comment faire ?

On suppose que l'on peut découper la forêt en 256 zones plus petites et de même forme.

- Quel est, en fonction de N, le nombre moyen m d'érables par zone ?
- Comment pensez-vous que l'on puisse donner une approximation de m en observant seulement un petit nombre de ces zones ?
- Comment choisir les zones à étudier ? Combien faut-il en prendre ?

Choix déterministe peut présenter des défauts (zones sélectionnées peuvent être non représentatives).
$\Rightarrow$ échantillonnage aléatoire : choisir aléatoirement (et avec remise) les zones à étudier.
Dans le cas le plus simple (arbres répartis de manière homogène), toutes les zones ont la même probabilité d'être observées.

Comment faire ?

On suppose que l'on peut découper la forêt en 256 zones plus petites et de même forme.

- Quel est, en fonction de N, le nombre moyen m d'érables par zone ?
- Comment pensez-vous que l'on puisse donner une approximation de m en observant seulement un petit nombre de ces zones ?
- Comment choisir les zones à étudier ? Combien faut-il en prendre ?

Choix déterministe peut présenter des défauts (zones sélectionnées peuvent être non représentatives).
$\Rightarrow$ échantillonnage aléatoire : choisir aléatoirement (et avec remise) les zones à étudier.
Dans le cas le plus simple (arbres répartis de manière homogène), toutes les zones ont la même probabilité d'être observées.

Comment faire ?

On suppose que l'on peut découper la forêt en 256 zones plus petites et de même forme.

- Quel est, en fonction de N, le nombre moyen m d'érables par zone ?
- Comment pensez-vous que l'on puisse donner une approximation de m en observant seulement un petit nombre de ces zones ?
- Comment choisir les zones à étudier ? Combien faut-il en prendre ?

Choix déterministe peut présenter des défauts (zones sélectionnées peuvent être non représentatives).

$\Rightarrow$ échantillonnage aléatoire : choisir aléatoirement (et avec remise) les zones à étudier.

Dans le cas le plus simple (arbres répartis de manière homogène), toutes les zones ont la même probabilité d'être observées.

Comment faire ?

On suppose que l'on peut découper la forêt en 256 zones plus petites et de même forme.

- Quel est, en fonction de N, le nombre moyen m d'érables par zone ?
- Comment pensez-vous que l'on puisse donner une approximation de m en observant seulement un petit nombre de ces zones ?
- Comment choisir les zones à étudier ? Combien faut-il en prendre ?

Choix déterministe peut présenter des défauts (zones sélectionnées peuvent être non représentatives).

$\Rightarrow$ échantillonnage aléatoire : choisir aléatoirement (et avec remise) les zones à étudier.

Dans le cas le plus simple (arbres répartis de manière homogène), toutes les zones ont la même probabilité d'être observées.

Comment faire ?

On suppose que l'on peut découper la forêt en 256 zones plus petites et de même forme.

- Quel est, en fonction de N, le nombre moyen m d'érables par zone ?
- Comment pensez-vous que l'on puisse donner une approximation de m en observant seulement un petit nombre de ces zones ?
- Comment choisir les zones à étudier ? Combien faut-il en prendre ?

Choix déterministe peut présenter des défauts (zones sélectionnées peuvent être non représentatives).
$\Rightarrow$ échantillonnage aléatoire : choisir aléatoirement (et avec remise) les zones à étudier.
Dans le cas le plus simple (arbres répartis de manière homogène), toutes les zones ont la même probabilité d'être observées.

Mise en pratique :

Exemples de choix aléatoires :

x	6	9	3	16	10	10	6	9	9	2	11	5	4	3	9	11
y	8	13	13	11	14	15	10	7	4	7	6	1	10	4	1	11
x	15	9	2	2	12	14	5	6	15	14	11	15	14	5	10	11
y	9	16	10	4	13	14	9	12	13	9	1	12	4	16	6	16
x	8	12	8	6	13	5	6	3								
y	10	8	15	1	1	7	10	4								

- Compter le nombre d'érables se trouvant dans les zones sélectionnées par cet échantillonnage aléatoire.
- Donner une approximation de m à partir des données recueillies.
- En déduire une approximation de N.

On peut aller plus loin :

- Calculer la variance de nos observations.
- Donner la valeur de l'intervalle de confiance associé à m.
- En déduire un intervalle de confiance pour N.

Mise en pratique :

Exemples de choix aléatoires :

x	6	9	3	16	10	10	6	9	9	2	11	5	4	3	9	11
y	8	13	13	11	14	15	10	7	4	7	6	1	10	4	1	11
x	15	9	2	2	12	14	5	6	15	14	11	15	14	5	10	11
y	9	16	10	4	13	14	9	12	13	9	1	12	4	16	6	16
x	8	12	8	6	13	5	6	3								
y	10	8	15	1	1	7	10	4								

- Compter le nombre d'érables se trouvant dans les zones sélectionnées par cet échantillonnage aléatoire.
- Donner une approximation de m à partir des données recueillies.
- En déduire une approximation de N.

On peut aller plus loin :

- Calculer la variance de nos observations.
- Donner la valeur de l'intervalle de confiance associé à m.
- En déduire un intervalle de confiance pour N.

Mise en pratique :

Exemples de choix aléatoires :

x	6	9	3	16	10	10	6	9	9	2	11	5	4	3	9	11
y	8	13	13	11	14	15	10	7	4	7	6	1	10	4	1	11
x	15	9	2	2	12	14	5	6	15	14	11	15	14	5	10	11
y	9	16	10	4	13	14	9	12	13	9	1	12	4	16	6	16
x	8	12	8	6	13	5	6	3								
y	10	8	15	1	1	7	10	4								

- Compter le nombre d'érables se trouvant dans les zones sélectionnées par cet échantillonnage aléatoire.
- Donner une approximation de m à partir des données recueillies.
- En déduire une approximation de N.

On peut aller plus loin :

- Calculer la variance de nos observations.
- Donner la valeur de l'intervalle de confiance associé à m.
- En déduire un intervalle de confiance pour N.

Mise en pratique :

Exemples de choix aléatoires :

x	6	9	3	16	10	10	6	9	9	2	11	5	4	3	9	11
y	8	13	13	11	14	15	10	7	4	7	6	1	10	4	1	11
x	15	9	2	2	12	14	5	6	15	14	11	15	14	5	10	11
y	9	16	10	4	13	14	9	12	13	9	1	12	4	16	6	16
x	8	12	8	6	13	5	6	3								
y	10	8	15	1	1	7	10	4								

- Compter le nombre d'érables se trouvant dans les zones sélectionnées par cet échantillonnage aléatoire.
- Donner une approximation de m à partir des données recueillies.
- En déduire une approximation de N.

On peut aller plus loin :

- Calculer la variance de nos observations.
- Donner la valeur de l'intervalle de confiance associé à m.
- En déduire un intervalle de confiance pour N.

Mise en pratique :

Exemples de choix aléatoires :

x	6	9	3	16	10	10	6	9	9	2	11	5	4	3	9	11
y	8	13	13	11	14	15	10	7	4	7	6	1	10	4	1	11
x	15	9	2	2	12	14	5	6	15	14	11	15	14	5	10	11
y	9	16	10	4	13	14	9	12	13	9	1	12	4	16	6	16
x	8	12	8	6	13	5	6	3								
y	10	8	15	1	1	7	10	4								

- Compter le nombre d'érables se trouvant dans les zones sélectionnées par cet échantillonnage aléatoire.
- Donner une approximation de m à partir des données recueillies.
- En déduire une approximation de N.

On peut aller plus loin :

- Calculer la variance de nos observations.
- Donner la valeur de l'intervalle de confiance associé à m.
- En déduire un intervalle de confiance pour N.

Mise en pratique :

Exemples de choix aléatoires :

x	6	9	3	16	10	10	6	9	9	2	11	5	4	3	9	11
y	8	13	13	11	14	15	10	7	4	7	6	1	10	4	1	11
x	15	9	2	2	12	14	5	6	15	14	11	15	14	5	10	11
y	9	16	10	4	13	14	9	12	13	9	1	12	4	16	6	16
x	8	12	8	6	13	5	6	3								
y	10	8	15	1	1	7	10	4								

- Compter le nombre d'érables se trouvant dans les zones sélectionnées par cet échantillonnage aléatoire.
- Donner une approximation de m à partir des données recueillies.
- En déduire une approximation de N.

On peut aller plus loin :

- Calculer la variance de nos observations.
- Donner la valeur de l'intervalle de confiance associé à m.
- En déduire un intervalle de confiance pour N.

Méthode de capture re-capture

Un autre problème concret

On souhaite donner un ordre de grandeur du nombre de poissons se trouvant dans une grande étendue d'eau.
On suppose que l'on ne peut observer qu'un poisson à la fois.

Expérimentation et mise en pratique :
-Vous disposez d'un récipient contenant un nombre N inconnu d'objets identiques.
-Vous ne pouvez sortir du récipient qu'un objet à la fois et devez mélanger avant chaque tirage.
-Vous avez à votre disposition un feutre indélébile.

- Comment pensez-vous que l'on puisse avoir une idée du nombre total d'objets présents dans la boîte ?
- A quoi peut servir le feutre ?

Un autre problème concret

On souhaite donner un ordre de grandeur du nombre de poissons se trouvant dans une grande étendue d'eau.
On suppose que l'on ne peut observer qu'un poisson à la fois.

Expérimentation et mise en pratique :
-Vous disposez d'un récipient contenant un nombre N inconnu d'objets identiques.
-Vous ne pouvez sortir du récipient qu'un objet à la fois et devez mélanger avant chaque tirage.
-Vous avez à votre disposition un feutre indélébile.

- Comment pensez-vous que l'on puisse avoir une idée du nombre total d'objets présents dans la boîte ?
- A quoi peut servir le feutre ?

Un autre problème concret

On souhaite donner un ordre de grandeur du nombre de poissons se trouvant dans une grande étendue d'eau.
On suppose que l'on ne peut observer qu'un poisson à la fois.

Expérimentation et mise en pratique :
-Vous disposez d'un récipient contenant un nombre N inconnu d'objets identiques.
-Vous ne pouvez sortir du récipient qu'un objet à la fois et devez mélanger avant chaque tirage.
-Vous avez à votre disposition un feutre indélébile.

- Comment pensez-vous que l'on puisse avoir une idée du nombre total d'objets présents dans la boîte ?
- A quoi peut servir le feutre ?

Un autre problème concret

On souhaite donner un ordre de grandeur du nombre de poissons se trouvant dans une grande étendue d'eau.
On suppose que l'on ne peut observer qu'un poisson à la fois.

Expérimentation et mise en pratique :
-Vous disposez d'un récipient contenant un nombre N inconnu d'objets identiques.
-Vous ne pouvez sortir du récipient qu'un objet à la fois et devez mélanger avant chaque tirage.
-Vous avez à votre disposition un feutre indélébile.

- Comment pensez-vous que l'on puisse avoir une idée du nombre total d'objets présents dans la boîte ?
- A quoi peut servir le feutre ?

Comment faire ?

Si un nombre connu N_M d'objets portent un signe distinctif.

- Quelle serait la probabilité P de choisir au hasard un objet portant un signe distinctif en fonction de N ?
- Supposons que l'on fasse n tirages avec remise dans le récipient en notant si l'objet porte une marque ou non. Comment peut-on donner une approximation p_n de P ?
- Déduire de p_n une approximation de N.

Revenons à notre problème concret.

- Proposer une méthode permettant de se ramener à l'étude d'un récipient contenant 20 objets marqués. [ETAPE DE CAPTURE]
- Mélanger les objets pour modéliser la répartition uniforme des individus marqués dans la zone d'étude. [DELAI D'ATTENTE]
- Effectuer ensuite $n = 40$ tirages avec remise. En déduire une approximation de N. [ETAPE DE RECAPTURE]

Comment faire ?

Si un nombre connu N_M d'objets portent un signe distinctif.

- Quelle serait la probabilité P de choisir au hasard un objet portant un signe distinctif en fonction de N ?
- Supposons que l'on fasse n tirages avec remise dans le récipient en notant si l'objet porte une marque ou non. Comment peut-on donner une approximation p_n de P ?
- Déduire de p_n une approximation de N.

Revenons à notre problème concret.

- Proposer une méthode permettant de se ramener à l'étude d'un récipient contenant 20 objets marqués. [ETAPE DE CAPTURE]
- Mélanger les objets pour modéliser la répartition uniforme des individus marqués dans la zone d'étude. [DELAI D'ATTENTE]
- Effectuer ensuite $n = 40$ tirages avec remise. En déduire une approximation de N. [ETAPE DE RECAPTURE]

Comment faire ?

Si un nombre connu N_M d'objets portent un signe distinctif.

- Quelle serait la probabilité P de choisir au hasard un objet portant un signe distinctif en fonction de N ?
- Supposons que l'on fasse n tirages avec remise dans le récipient en notant si l'objet porte une marque ou non. Comment peut-on donner une approximation p_n de P ?
- Déduire de p_n une approximation de N.

Revenons à notre problème concret.

- Proposer une méthode permettant de se ramener à l'étude d'un récipient contenant 20 objets marqués. [ETAPE DE CAPTURE]
- Mélanger les objets pour modéliser la répartition uniforme des individus marqués dans la zone d'étude. [DELAI D'ATTENTE]
- Effectuer ensuite $n = 40$ tirages avec remise. En déduire une approximation de N. [ETAPE DE RECAPTURE]

Comment faire ?

Si un nombre connu N_M d'objets portent un signe distinctif.

- Quelle serait la probabilité P de choisir au hasard un objet portant un signe distinctif en fonction de N ?
- Supposons que l'on fasse n tirages avec remise dans le récipient en notant si l'objet porte une marque ou non. Comment peut-on donner une approximation p_n de P ?
- Déduire de p_n une approximation de N.

Revenons à notre problème concret.

- Proposer une méthode permettant de se ramener à l'étude d'un récipient contenant 20 objets marqués. [ETAPE DE CAPTURE]
- Mélanger les objets pour modéliser la répartition uniforme des individus marqués dans la zone d'étude. [DELAI D'ATTENTE]
- Effectuer ensuite $n = 40$ tirages avec remise. En déduire une approximation de N. [ETAPE DE RECAPTURE]

Comment faire ?

Si un nombre connu N_M d'objets portent un signe distinctif.

- Quelle serait la probabilité P de choisir au hasard un objet portant un signe distinctif en fonction de N ?
- Supposons que l'on fasse n tirages avec remise dans le récipient en notant si l'objet porte une marque ou non. Comment peut-on donner une approximation p_n de P ?
- Déduire de p_n une approximation de N.

Revenons à notre problème concret.

- Proposer une méthode permettant de se ramener à l'étude d'un récipient contenant 20 objets marqués.
  [ETAPE DE CAPTURE]
- Mélanger les objets pour modéliser la répartition uniforme des individus marqués dans la zone d'étude.
  [DELAI D'ATTENTE]
- Effectuer ensuite $n = 40$ tirages avec remise. En déduire une approximation de N. [ETAPE DE RECAPTURE]

Comment faire ?

Si un nombre connu N_M d'objets portent un signe distinctif.

- Quelle serait la probabilité P de choisir au hasard un objet portant un signe distinctif en fonction de N ?
- Supposons que l'on fasse n tirages avec remise dans le récipient en notant si l'objet porte une marque ou non. Comment peut-on donner une approximation p_n de P ?
- Déduire de p_n une approximation de N.

Revenons à notre problème concret.

- Proposer une méthode permettant de se ramener à l'étude d'un récipient contenant 20 objets marqués. [ETAPE DE CAPTURE]
- Mélanger les objets pour modéliser la répartition uniforme des individus marqués dans la zone d'étude. [DELAI D'ATTENTE]
- Effectuer ensuite $n = 40$ tirages avec remise. En déduire une approximation de N. [ETAPE DE RECAPTURE]

Pour aller un peu plus loin

Supposons que sur 1000 observations on ait une proportion de 0.199 poissons marqués.

- Donner un intervalle de confiance pour P.
- En déduire un intervalle de confiance pour N.

Méthode du maximum de vraisemblance

Un autre exemple concret

L'idée principale est de choisir la valeur de N avec laquelle on a le plus de chance d'obtenir les observations qui ont été faites.

On cherche à donner une approximation du nombre de rhinocéros se trouvant dans une zone géographique assez vaste.

On suppose que les rhinocéros ont

- une probabilité de $\frac{1}{3}$ d'aller se désaltérer à l'un des rares points d'eau en fin d'après-midi.
- des comportements indépendants les uns des autres.

On fait les observations suivantes :

Nombre de rhinocéros	0	1	2	2	1

Un autre exemple concret

L'idée principale est de choisir la valeur de N avec laquelle on a le plus de chance d'obtenir les observations qui ont été faites.

On cherche à donner une approximation du nombre de rhinocéros se trouvant dans une zone géographique assez vaste.

On suppose que les rhinocéros ont

- une probabilité de $\frac{1}{3}$ d'aller se désaltérer à l'un des rares points d'eau en fin d'après-midi.
- des comportements indépendants les uns des autres.

On fait les observations suivantes :

Nombre de rhinocéros	0	1	2	2	1

Un autre exemple concret

L'idée principale est de choisir la valeur de N avec laquelle on a le plus de chance d'obtenir les observations qui ont été faites.

On cherche à donner une approximation du nombre de rhinocéros se trouvant dans une zone géographique assez vaste.

On suppose que les rhinocéros ont

- une probabilité de $\frac{1}{3}$ d'aller se désaltérer à l'un des rares points d'eau en fin d'après-midi.
- des comportements indépendants les uns des autres.

On fait les observations suivantes :

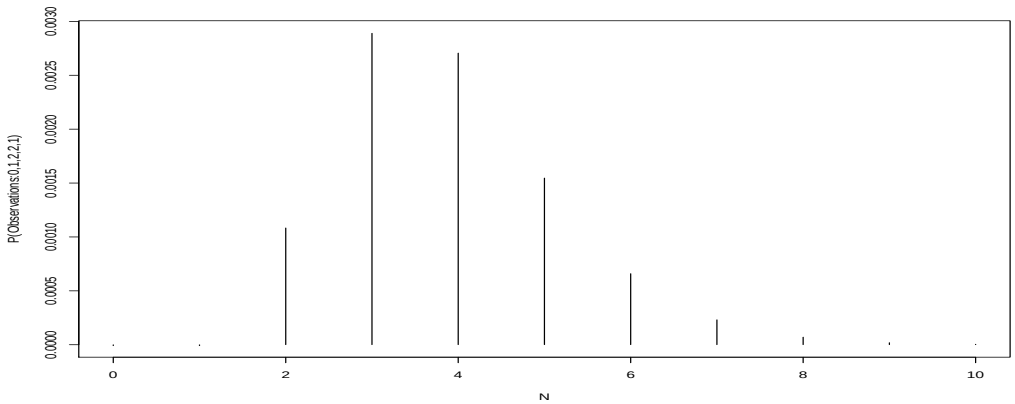
Nombre de rhinocéros	0	1	2	2	1

La méthode du maximum de vraisemblance

On peut donner une expression explicite de la probabilité d'avoir observé $0, 1, 2, 2, 1$ rhinocéros en fonction de N :

$$P_N(\textit{Observations} : 0, 1, 2, 2, 1) = \left(\frac{2}{3}\right)^{5N} \frac{N^4(N-1)^2}{2^8}.$$

- Est-ce que N peut valoir 0 ou 1 ?
- A partir du graphique, pour quelle valeur de N la probabilité est la plus forte que se produise ce que l'on a observé ?



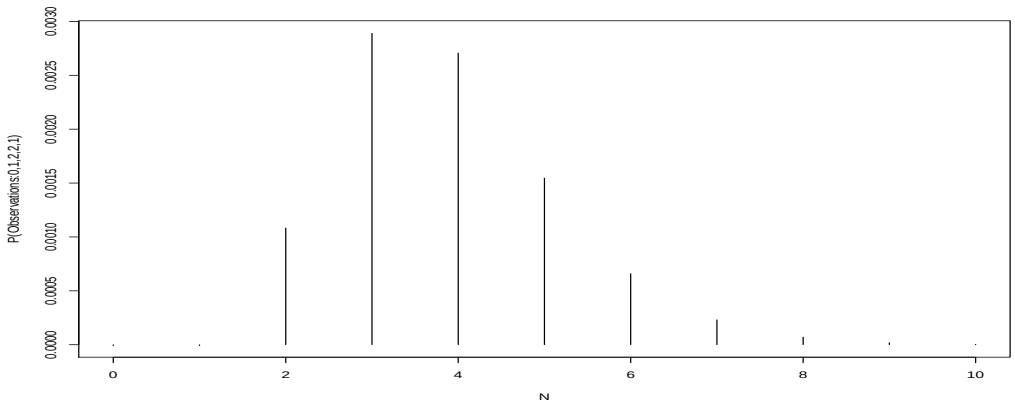
- En déduire la valeur de N la plus vraisemblable.

La méthode du maximum de vraisemblance

On peut donner une expression explicite de la probabilité d'avoir observé $0, 1, 2, 2, 1$ rhinocéros en fonction de N :

$$P_N(\textit{Observations} : 0, 1, 2, 2, 1) = \left(\frac{2}{3}\right)^{5N} \frac{N^4(N-1)^2}{2^8}.$$

- Est-ce que N peut valoir 0 ou 1 ?
- A partir du graphique, pour quelle valeur de N la probabilité est la plus forte que se produise ce que l'on a observé ?



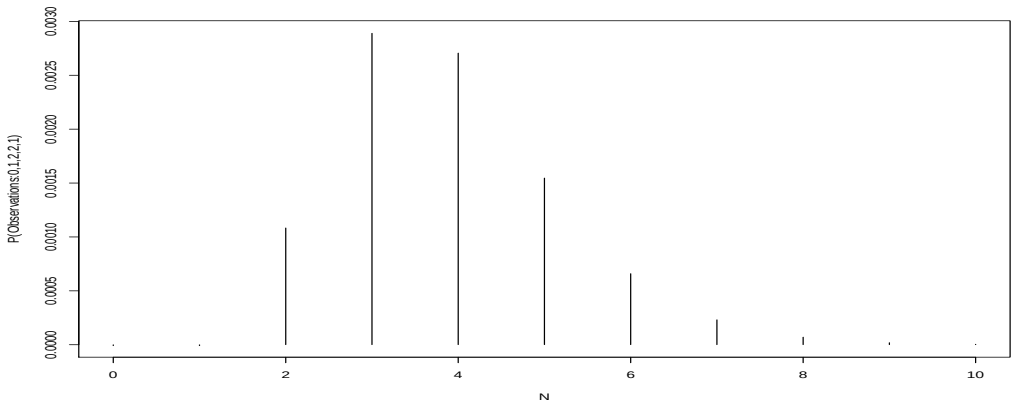
- En déduire la valeur de N la plus vraisemblable.

La méthode du maximum de vraisemblance

On peut donner une expression explicite de la probabilité d'avoir observé $0, 1, 2, 2, 1$ rhinocéros en fonction de N :

$$P_N(\textit{Observations} : 0, 1, 2, 2, 1) = \left(\frac{2}{3}\right)^{5N} \frac{N^4(N-1)^2}{2^8}.$$

- Est-ce que N peut valoir 0 ou 1 ?
- A partir du graphique, pour quelle valeur de N la probabilité est la plus forte que se produise ce que l'on a observé ?



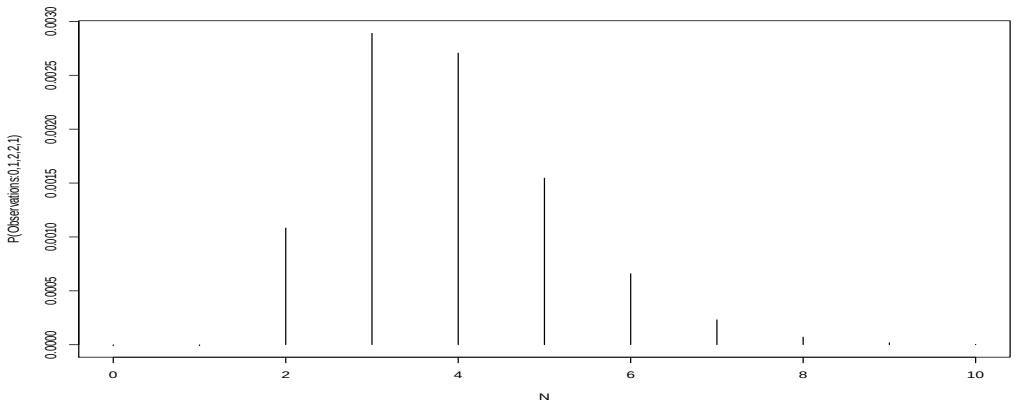
- En déduire la valeur de N la plus vraisemblable.

La méthode du maximum de vraisemblance

On peut donner une expression explicite de la probabilité d'avoir observé $0, 1, 2, 2, 1$ rhinocéros en fonction de N :

$$P_N(\textit{Observations} : 0, 1, 2, 2, 1) = \left(\frac{2}{3}\right)^{5N} \frac{N^4(N-1)^2}{2^8}.$$

- Est-ce que N peut valoir 0 ou 1 ?
- A partir du graphique, pour quelle valeur de N la probabilité est la plus forte que se produise ce que l'on a observé ?



- En déduire la valeur de N la plus vraisemblable.

Un autre exemple :

Imaginons par exemple qu'un nouveau participant arrive en fin de séance et qu'il ne sache pas combien de dés vous avez lancés au début.

S'il dispose uniquement du nombre de 5 ou 6 observés pour chaque lancé de 5 dés, il peut utiliser la méthode du maximum de vraisemblance pour déterminer à partir des observations la valeur la plus vraisemblable de dés que vous avez jetés.

En effet, ici aussi le nombre de 5 ou 6 obtenus à chacun de vos lancés peut être modélisé par une variable de loi $Bin(N, \frac{1}{3})$, où N représente le nombre inconnu de dés que vous avez lancés.

Conclusion

Conclusion et commentaires

- méthodes statistiques permettent de donner une approximation du nombre d'individus mais aussi de donner un ordre de grandeur de la précision de notre estimation.
- certaine variabilité de nos résultats, qui diminue cependant lorsque le nombre d'observations augmente.
- important de prendre en compte la qualité des données (taille de l'échantillon, représentativité) lorsque l'on cherche à interpréter les résultats donnés par une méthode d'estimation (par exemple un sondage).
- méthodes plus fines et précises sont généralement utilisées.

Conclusion et commentaires

- méthodes statistiques permettent de donner une approximation du nombre d'individus mais aussi de donner un ordre de grandeur de la précision de notre estimation.
- certaine variabilité de nos résultats, qui diminue cependant lorsque le nombre d'observations augmente.
- important de prendre en compte la qualité des données (taille de l'échantillon, représentativité) lorsque l'on cherche à interpréter les résultats donnés par une méthode d'estimation (par exemple un sondage).
- méthodes plus fines et précises sont généralement utilisées.

Conclusion et commentaires

- méthodes statistiques permettent de donner une approximation du nombre d'individus mais aussi de donner un ordre de grandeur de la précision de notre estimation.
- certaine variabilité de nos résultats, qui diminue cependant lorsque le nombre d'observations augmente.
- important de prendre en compte la qualité des données (taille de l'échantillon, représentativité) lorsque l'on cherche à interpréter les résultats donnés par une méthode d'estimation (par exemple un sondage).
- méthodes plus fines et précises sont généralement utilisées.

Conclusion et commentaires

- méthodes statistiques permettent de donner une approximation du nombre d'individus mais aussi de donner un ordre de grandeur de la précision de notre estimation.
- certaine variabilité de nos résultats, qui diminue cependant lorsque le nombre d'observations augmente.
- important de prendre en compte la qualité des données (taille de l'échantillon, représentativité) lorsque l'on cherche à interpréter les résultats donnés par une méthode d'estimation (par exemple un sondage).
- méthodes plus fines et précises sont généralement utilisées.

Tests médicaux

Les données

On souhaite étudier l'efficacité d'un test médical permettant de dépister une maladie présente chez 0.5% de la population. Le fabricant de ce test indique qu'il est positif pour 99% des malades et négatif pour 95% des personnes saines.
bigskip

- On dispose de données (simulées) concernant un échantillon de 1000000 individus.

- Pour chacun de ces individus on observe une variable Santé donnant l'état de santé de l'individu et une variable Test donnant le résultat du test de dépistage.

Sondages

Les données

Une enquête est menée auprès d'un échantillon de n électeurs potentiels. On leur demande s'ils sont plutôt favorables au candidat 1 ou au candidat 2. On obtient 53% d'intentions de vote pour le candidat numéro 1.
Peut-on affirmer que le candidat 1 rassemble plus d'intentions de vote dans l'ensemble de la population ?

Quelle information nous manque ?

On suppose que n=100, l'intervalle de confiance pour p est [0.4321784; 0.6278216]. Peut-on conclure quelque chose ?

Quelle taille d'échantillon est nécessaire pour avoir une précision de 0.03 ?

Les données

Une enquête est menée auprès d'un échantillon de n électeurs potentiels. On leur demande s'ils sont plutôt favorables au candidat 1 ou au candidat 2. On obtient 53% d'intentions de vote pour le candidat numéro 1.
Peut-on affirmer que le candidat 1 rassemble plus d'intentions de vote dans l'ensemble de la population ?

Quelle information nous manque ?

On suppose que n=100, l'intervalle de confiance pour p est [0.4321784; 0.6278216]. Peut-on conclure quelque chose ?

Quelle taille d'échantillon est nécessaire pour avoir une précision de 0.03 ?

Les données

Une enquête est menée auprès d'un échantillon de n électeurs potentiels. On leur demande s'ils sont plutôt favorables au candidat 1 ou au candidat 2. On obtient 53% d'intentions de vote pour le candidat numéro 1.
Peut-on affirmer que le candidat 1 rassemble plus d'intentions de vote dans l'ensemble de la population ?

Quelle information nous manque ?

On suppose que n=100, l'intervalle de confiance pour p est $[0.4321784; 0.6278216]$. Peut-on conclure quelque chose ?

Quelle taille d'échantillon est nécessaire pour avoir une précision de 0.03 ?

Sondages sur données sensibles

La méthode utilisée :

On souhaite évaluer la proportion P de personnes faisant des téléchargements illégaux dans une population.
On ne peux pas leur poser directement la question car ils risquent de ne pas nous dire la vérité par peur d'éventuels problèmes judiciaires.
Un statisticien propose la démarche suivante : demander aux individus de lancer un dé dans un isoloir puis de donner seulement la reponse à la question :

- Q1 : avez-vous obtenu 1 ou 2 ?, s'ils telechargent illegalement.
- Q2 : avez-vous obtenu 3,4,5 ou 6 ?, s'ils ne telechargent pas illegalement.

Seule la personne interrogée connait la question à laquelle elle répond.

La méthode utilisée :

On souhaite évaluer la proportion P de personnes faisant des téléchargements illégaux dans une population.
On ne peux pas leur poser directement la question car ils risquent de ne pas nous dire la vérité par peur d'éventuels problèmes judiciaires.
Un statisticien propose la démarche suivante : demander aux individus de lancer un dé dans un isoloir puis de donner seulement la reponse à la question :

- Q1 : avez-vous obtenu 1 ou 2 ?, s'ils telechargent illegalement.
- Q2 : avez-vous obtenu 3,4,5 ou 6 ?, s'ils ne telechargent pas illegalement.

Seule la personne interrogée connait la question à laquelle elle répond.

Comment marche-t-elle ?

- Quelle est la probabilité qu'une personne qui télécharge illégalement réponde OUI ?
- Quelle est la probabilité qu'une personne qui ne télécharge pas illégalement réponde OUI ?
- Quelle est la probabilité qu'une personne interrogée réponde OUI ?

$$p_{OUI} = p_{PIRATES}\frac{1}{3} + (1 - p_{PIRATES})\frac{2}{3} = \frac{2 - p_{PIRATES}}{3}.$$

Comment marche-t-elle ?

- Quelle est la probabilité qu'une personne qui télécharge illégalement réponde OUI ?
- Quelle est la probabilité qu'une personne qui ne télécharge pas illégalement réponde OUI ?
- Quelle est la probabilité qu'une personne interrogée réponde OUI ?

$$p_{OUI} = p_{PIRATES}\frac{1}{3} + (1 - p_{PIRATES})\frac{2}{3} = \frac{2 - p_{PIRATES}}{3}.$$

Comment marche-t-elle ?

- Quelle est la probabilité qu'une personne qui télécharge illégalement réponde OUI ?
- Quelle est la probabilité qu'une personne qui ne télécharge pas illégalement réponde OUI ?
- Quelle est la probabilité qu'une personne interrogée réponde OUI ?

$$p_{OUI} = p_{PIRATES}\frac{1}{3} + (1 - p_{PIRATES})\frac{2}{3} = \frac{2 - p_{PIRATES}}{3}.$$

Comment marche-t-elle ?

- Quelle est la probabilité qu'une personne qui télécharge illégalement réponde OUI ?
- Quelle est la probabilité qu'une personne qui ne télécharge pas illégalement réponde OUI ?
- Quelle est la probabilité qu'une personne interrogée réponde OUI ?

$$p_{OUI} = p_{PIRATES}\frac{1}{3} + (1 - p_{PIRATES})\frac{2}{3} = \frac{2 - p_{PIRATES}}{3}.$$

Comment marche-t-elle ?

- Comment peut-on donner une valeur approchée de p_{OUI} ?
- En déduire une valeur approchée de $p_{PIRATES}$.

$$p_{PIRATES} = 2 - 3p_{OUI} \Rightarrow \hat{p}_{PIRATES} = 2 - 3\hat{p}_{OUI}$$

- Application.

Comment marche-t-elle ?

- Comment peut-on donner une valeur approchée de p_{OUI} ?
- En déduire une valeur approchée de $p_{PIRATES}$.

$$p_{PIRATES} = 2 - 3p_{OUI} \Rightarrow \hat{p}_{PIRATES} = 2 - 3\hat{p}_{OUI}$$

- Application.

Comment marche-t-elle ?

- Comment peut-on donner une valeur approchée de p_{OUI} ?
- En déduire une valeur approchée de $p_{PIRATES}$.

$$p_{PIRATES} = 2 - 3p_{OUI} \Rightarrow \hat{p}_{PIRATES} = 2 - 3\hat{p}_{OUI}$$

- Application.