Scénario 9

ESTIMATION D'UN PARAMETRE

1 Présentation générale

1.1 Objectifs

- expliquer la différence entre un paramètre calculé sur les données d'une expérience et le paramètre de référence correspondant.
- donner un intervalle de confiance d'un paramètre dans un cas simple.

1.2 Les étapes

1. Explication des objectifs du scénario : ce qu'on sait déjà.
2. Estimation de la proportion de haricots rouges (ou d'autre choses) à partir de tirages dans un grand sac.
3. Propriétés d'un estimateur. Une histoire de taxis (*facultatif*).
4. Intervalle de confiance : on donne une estimation avec une marge d'erreur; on précise cette erreur.
5. Utilisation d'abaques pour construire un intervalle de confiance d'une proportion.
6. Comment sont construites ces abaques ou aperçu de la loi binomiale (*facultatif*)
7. Evaluation orale collective.

1.3 Matériel

- pour les stagiaires : 2 feutres de couleur
- pour le formateur : un grand tableau et un sac contenant au choix
  - un mélange de haricots blancs et rouges
  - des pistaches ouvertes et fermées
  - des billes de deux sortes
  - des graines ...

1.4 La forme

Travail essentiellement collectif, avec écriture au tableau.
Feuilles pour les résultats, avec les tables et les abaques distribuées aux stagiaires pour annotation au fur et à mesure.

1.5 Evaluation

Orale et collective sur un deuxième exemple : estimation de la médiane d'une mesure dans une population.

1.6 Prérequis

Population
Echantillon
Calcul d'une médiane
Utilisation de graphiques : lecture de l'ordonnée d'un point sur une courbe pour une abscisse donnée.

1.7 Durée

Une séance

2 Les étapes

2.1 Etape 1 : Explication des objectifs du scénario

But de cette étape : identifier les préacquis.
Enoncer les deux objectifs et faire exprimer comment les stagiaires les comprennent. Faire préciser

- ce qu'est un paramètre (c'est à dire un paramètre résumant une variable, ou la distribution de cette variable);
- pourquoi il y a une différence entre deux types de paramètres;
- identifier la différence entre une population et les données (échantillon).
  Si besoin, suggérer un exemple de population : les agents INRA, le sac de haricots (ou de pistaches, ou de billes...) qu'on a apporté et la variable d'étude (le salaire, le grade ou l'âge, la couleur);
- un intervalle de confiance (c'est une réponse plus ou moins une erreur).

2.2 Etape 2 : Estimation de la proportion de haricots rouges

Le paramètre à estimer est la proportion inconnue de haricots rouges notée p. On choisit une taille d'échantillon (10 ou 20 pour des calculs rapides et pour utiliser la table simple). On choisit la méthode d'échantillonnage : avec ou sans remise.

Dire que ce choix impliquera des calculs différents pour l'intervalle de confiance, que les plus classiques sont ceux du tirage avec remise, qu'on adopte donc. Mais si le nombre de haricots est grand et p pas trop près de 0 ou 1, cela revient au même. On peut donc gagner du temps pour la manipulation.

Chaque stagiaire tire un échantillon et note la proportion observée.
Sur la feuille stagiaire 1 et au tableau, on porte les proportions successives observées.
On remarque la variabilité des résultats : certains sont très faux, presque tous sont faux puisqu'un seul au plus est la vraie valeur p.

La proportion observée est aléatoire : elle dépend de l'expérience (du tirage particulier). Chaque proportion observée est une **estimation** de la proportion inconnue.

2.3 Etape 3 : (*facultative*) Propriétés d'un estimateur

Dans l'exemple des haricots, l'estimation de la proportion inconnue par la proportion observée est naturelle. On appelle estimateur la formule qui calcule ainsi une estimation à partir de chaque échantillon observé. On le note $\hat{p}$ et $\hat{p}_1$, $\hat{p}_2$,... $\hat{p}_{10}$ les valeurs successives obtenues de cet estimateur.

Il y a des cas ou le choix de l'estimateur n'est pas simple.

Exemple des taxis (d'après Nœther, 1976)
Dans l'attente d'un taxi vide, vous notez les n° des taxis pleins qui passent et vous obtenez :

405	280	73	440	179

Question : Quelle estimation faites-vous du nombre de taxis de la ville?

Réponse : Il y a plusieurs réponses possibles :

1. $440 + 72 = 512$
   On ajoute autant au-dessus du plus grand numéro qu'il y en a au-dessous du plus petit, c'est à dire 72, si la numérotation commence à 1.

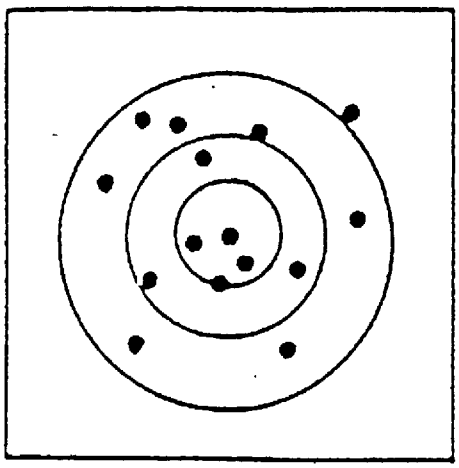
2. $280 + 279 = 559$
   La médiane observée est 280 : si c'était la vraie médiane, on aurait autant de numéros à droite qu'à gauche de 280.

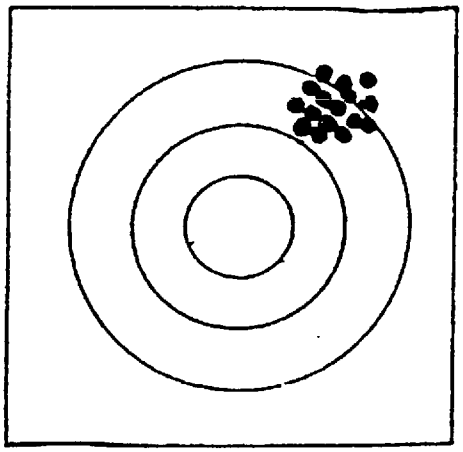
3. On fait la moyenne des intervalles entre deux numéros successifs en ajoutant le premier intervalle de 72 et on ajoute ce nombre à 440.

Quand on a le choix entre plusieurs estimateurs, comment choisir le meilleur? Que veut dire "meilleur"?

- tomber autour de la valeur cherchée, être **en moyenne** égal à cette valeur (en moyenne sur un nombre infini d'expériences). C'est la propriété "*sans biais*".
- ne pas trop varier d'une expérience à l'autre, avoir une petite variance. C'est la propriété "*de variance minimale*".

Faire les deux dessins suivants au tableau.





Quelle propriété correspond à chacun d'eux?
Laquelle préférer?
Si on peut avoir un estimateur sans biais et de variance minimum, c'est le mieux. Mais on ne connaît pas toujours un tel estimateur.
C'est l'objet de la statistique et le travail des statisticiens d'étudier ces problèmes.

2.4 Etape 4 : Intervalle de confiance

Que faire puisqu'on n'est pas sûr de la réponse?

Réponses possibles :

1. faire la moyenne de toutes les estimations obtenues (*montrer que cela revient à changer la taille de l'échantillon, qu'intuitivement cela doit être meilleur, mais que c'est encore aléatoire*).
2. donner une marge d'erreur. Comment obtenir cette marge? *Ecouter les propositions, les noter au tableau* (encadrer par max et min, estimer une variance ou écart...).

Donner la feuille stagiaire 2 avec la table de l'intervalle de confiance (*sans dire pourquoi 95% pour l'instant*) et complèter la feuille stagiaire 1 et au tableau.

Remarquer que ces intervalles sont variables, eux aussi. Constater peut-être que certains ont une intersection vide, sinon dire que cela peut arriver.
Expliquer que si on faisait un très grand (infini) nombre d'expériences, plus de 95% des intervalles obtenus seraient corrects, c'est à dire contiendraient la vraie valeur.

Dévoiler maintenant la valeur inconnue (vous aurez compté vos haricots avant!).
Tracer une verticale sur les intervalles dessinés (sur la feuille stagiaire 1) et remplir la colonne erreur avec des oui ou des non (ou des 0 et des 1).
Refaire éventuellement quelques échantillons pour convaincre.

Montrer la simulation d'une binomiale B (10, 0.2) (feuille stagiaire 4) en disant que c'est le cas où la proportion est p = 0,2.
Coefficient de confiance (ou niveau de confiance) = probabilité que, pour une expérience quelconque, l'intervalle calculé contienne la valeur p inconnue. Ici ce coefficient est $\geq$ 0,95.

Question : Quelle est la probabilité de donner un intervalle qui ne contient pas p? (c'est la **probabilité d'erreur**)
Si je veux diminuer cette probabilité d'erreur que faire?

Réponse : agrandir les intervalles.
Eventuellement donner la feuille stagiaire 2bis et sur la feuille stagiaire 1 agrandir (avec un feutre de couleur) les intervalles et modifier la colonne erreur.

Remarque importante : pour une expérience particulière, l'intervalle obtenu est juste ou faux : malheureusement on n'en sait rien!

2.5 Etape 5 : utilisation des abaques

Distribuer la feuille stagiaire 3. Reproduire schématiquement au tableau le cadre et les deux courbes correspondant à la taille d'échantillon choisie.
Montrer qu'on trace la verticale d'abscisse $\hat{p}$ et qu'on obtient l'intervalle avec les intersection entre les deux courbes.
Faire retrouver un ou quelques intervalles à chaque stagiaire.

2.6 Etape 6 : (*facultative*)Construction des abaques. Aperçu de la loi binomiale

Les abaques se lisent en traçant des lignes verticales, se construisent en traçant des lignes horizontales.

On peut introduire les choses ainsi :
La proportion p de haricots rouges peut s'interprèter comme une **probabilité**. *Obtenir la formulation correcte.*

Envisager le lancer d'un dé à 10 faces. On note s'il s'agit d'un chiffre ≤ 2 ($\Rightarrow$ p = 0,2).

Envisager le lancer de 10 dés. On compte le nombre de chiffres ≤ 2.
Faire la relation avec les haricots : si la proportion de rouges est 0,2, les deux expériences sont comparables.

Pour étudier la distribution de $\hat{p}$ ou plutôt de $N = 10\hat{p}$, on peut faire une simulation avec 10 dés.
Envisager le début de cette simulation : commencer un stem-and-leaf. Regarder celui obtenu (feuille stagiaire 4).
Expliquer qu'il se régularise si le nombre de simulations augmente.

Distribuer la feuille stagiaire 5.
Dire qu'il s'agit de la distribution binomiale B(10,1/5).
Interprèter les probabilités du tableau.

Lecture des abaques dans l'autre sens : une idée de comment on peut construire les abaques.

Pour p = 0,2 , trouver un intervalle dans lequel N soit dans au moins 95% des cas : on additionne les probabilités de la queue de droite :
0 + 0 + 0,0001 + 0,0008 + 0,0055 + 0,0264 = 0,0328
si on ajoute Pr { N = 4 } on dépasse 0,05 pour la queue de droite, donc il faut choisir l'intervalle [0 5]

Voir sur l'abaque : pour la ligne horizontale d'ordonnée 0,2 on lit [0 5,2]. C'est la même chose, puisque N ne prend que des valeurs entières.

Pour p = 0,5 par lecture sur l'abaque on obtient l'intervalle [1,4 8,4].
Par calcul à partir de la table (feuille 5) :

$$(0{,}0010 + 0{,}0098) \times 2 = 0{,}0216$$
$$(0{,}0010 + 0{,}0098 + 0{,}0440) \times 2 = 0{,}1096$$

on dépasse 0,05.

L'intervalle pour les valeurs de N est [2 8]. C'est la même chose que N dans [1,4 8,4].

Résumé : par simulation ou par calcul théorique, à partir de chaque valeur de p théorique, on peut obtenir les deux valeurs entre lesquelles N se trouve dans 95% des cas. On les note sur le graphique et on fait cela pour toutes les valeurs de p : les deux courbes correspondant à la taille choisie de l'échantillon se construisent en gros ainsi.
Attention : dans cette étape on calcule des intervalles pour N ou pour $\hat{p}$ dont les bornes sont fixées, alors que l'intervalle de confiance est un intervalle pour p, dont les bornes sont aléatoires. *Ce sera probablement l'occasion de confusions. Ne pas insister.*

2.7 Evaluation collective

- Le problème est de connaître la médiane des salaires (ou des âges) des chefs de service à l'INRA. Le service du personnel refuserait de donner l'information!

  A partir des salaires (un peu inventés si besoin) des chefs des stagiaires, on va essayer d'estimer ce **salaire médian** noté s.

- Demander des solutions.

  Vérifier si la distinction salaire médian des chefs de l'institut, salaire médian des chefs des stagiaires (population, échantillon) est acquise.

  Si pas d'idées, faire réordonner les valeurs de l'échantillon et demander si $\hat{s}$ est la valeur cherchée. Ce qu'elle serait avec un autre groupe de stagiaires.

- Demander comment on peut trouver un intervalle.

  Si besoin, suggérer [min , max], avec les valeurs maximale et minimale de l'échantillon.

  Demander ce qu'il faut connaître pour utiliser cet intervalle. Obtenir la phrase correcte concernant la probabilité de contenir s.

- En utilisant la table donnée en annexe, donner le coefficient de confiance de [min , max].

  Demander ce qui se passe si on diminue l'intervalle (ou si on veut diminuer ce coefficient). Donner les autres valeurs. Choisir un intervalle de coefficient $\geq 0{,}95$.

2.8 Mots-clés

population
échantillon
variable aléatoire
paramètre théorique
estimation du paramètre
estimation
estimateur
toutes les expériences possibles
l'expérience particulière dont on a les données

2.9 Vocabulaire

population (scénario 3)
échantillon (scénario 3)
paramètre
estimateur
intervalle de confiance
coefficient d'un intervalle de confiance
probabilité d'erreur

2.10 Références bibliographiques

CERESTA (1973) *Tables statistiques.* Revue de statistique appliquée.

LEFORT G. (1967) *Mathématiques pour les sciences biologiques et agronomiques.* Armand Colin.

NEMENYI P., DIXON S.K., WHITE N.B.Jr, HEDSTROM M. (1977) *Statistics from scratch.* Holden Day.

NŒTHER G. (1976) *Introduction to statistics, a nonparametric approach.* Houghton. Mifflin.

PEARSON E.S., HARTEY (H.O.) (1970) *Biometrika tables for statisticians.* Cambridge University Press.

3 Annexes

3.1 Documents stagiaires

Présentation du scénario
Feuilles stagiaires 1, 2, 2bis, 3, 4, 5
Vocabulaire du scénario 9

3.2 Documents formateur

5 feuilles dont :
Table de l'intervalle d'une proportion
Abaque de coefficient 0,99 de l'estimation d'une proportion
Table de l'intervalle de confiance de la médiane