Travail de l'élève 1 : En Syldavie, Norbert fait des études statistiques sur le daltonisme des éléphants syldaves pratiquant le saut en parachute.
Il pense que 46% d'entre eux sont des mâles et 18% des éléphants ont plus de 60 ans.
Il a prélevé un échantillon de 400 éléphants parmi lesquels 195 sont des mâles et 313 ont moins de 60 ans.
L'échantillon a été réalisé par un tirage au hasard et il peut être assimilé à un tirage avec remise.

1. L'échantillon de Norbert est-il représentatif ?
2. L'étude de Norbert montre que dans cet échantillon, 29% des éléphants sont daltoniens.
   Estimer la proportion d'éléphants daltoniens dans la population totale.

Travail de l'élève 2 : On reprend le contexte de l'activité précédente.
Norbert veut effectuer une étude plus précise avec un échantillon plus grand.
Il a un échantillon de 1200 éléphants parmi lesquels 562 sont des mâles et 951 ont moins de 60 ans.

1. **a.** Pourquoi aimeront-on trouver une autre méthode que celles employées précédemment pour savoir si l'échantillon de Norbert est représentatif ?
   b. Quel nouvel outil de Terminale pourrait-on utiliser ?
   c. Proposer un nouvel intervalle de fluctuation au seuil de 95%.
   Quelle différence a-t-il avec les précédents ? Quels sont ses avantages ? ses inconvénients ?
2. L'étude de Norbert montre que dans cet échantillon, 32% des éléphants sont daltoniens.
   Estimer la proportion d'éléphants daltoniens dans la population totale.

Conclusion

Soit p la probabilité d'apparition d'un caractère dans une population totale.

– **Lorsque l'on connait** p, ou l'on fait une hypothèse sur la valeur de p, on utilise un **intervalle de .... à** 95% pour estimer la fréquence d'apparition f du caractère dans un échantillon de taille n.

Si $\qquad$ $\mathrm{I} = \left[\ldots\ldots\ldots\ldots\ldots\ldots\ldots ; \ldots\ldots\ldots\ldots\ldots\ldots\ldots\right]$

ou $\qquad$ $\mathrm{I}' = \left[\ldots\ldots\ldots ; \ldots\ldots\ldots\right]$

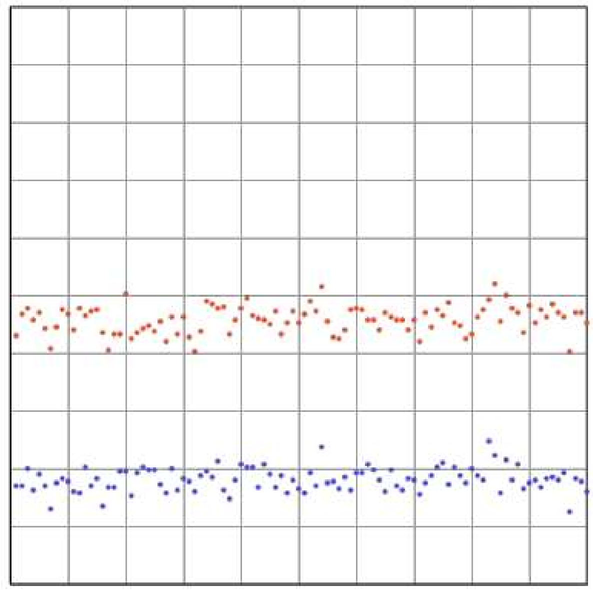
où pour $\mathrm{X} \hookrightarrow$ .... et :
- a désigne le plus petit entier k à partir duquel $\mathrm{P}(\mathrm{X} \leq k)$ dépasse ....
- b désigne le plus petit entier k à partir duquel $\mathrm{P}(\mathrm{X} \leq k)$ dépasse ....

– **Lorsque l'on ne connait pas** p, on utilise un **intervalle de .... à** 95% pour estimer la valeur de p à partir de la fréquence f d'apparition du caractère dans un échantillon de taille n.

Si $\qquad$ $\mathrm{J} = \left[\ldots\ldots\ldots\ldots\ldots\ldots\ldots ; \ldots\ldots\ldots\ldots\ldots\ldots\ldots\right]$

Question 1 : L'échantillon de Norbert est-il représentatif ?

```
VARIABLES
  a EST_DU_TYPE NOMBRE
  Males EST_DU_TYPE NOMBRE
  Vieux EST_DU_TYPE NOMBRE
  compteur EST_DU_TYPE NOMBRE
  compteur2 EST_DU_TYPE NOMBRE
DEBUT_ALGORITHME
  POUR compteur ALLANT_DE 1 A 100
    DEBUT_POUR
    Males PREND_LA_VALEUR 0
    Vieux PREND_LA_VALEUR 0
    POUR compteur2 ALLANT_DE 1 A 400
      DEBUT_POUR
      a PREND_LA_VALEUR random()
      SI (a<=0.46) ALORS
        DEBUT_SI
        Males PREND_LA_VALEUR Males+1
        FIN_SI
      SI (a<=0.18) ALORS
        DEBUT_SI
        Vieux PREND_LA_VALEUR Vieux+1
        FIN_SI
      FIN_POUR
    TRACER_POINT (compteur,Males/400)
    TRACER_POINT (compteur,Vieux/400)
    FIN_POUR
FIN_ALGORITHME
```



Xmin: 0 ; Xmax: 100 ; Ymin: 0 ; Ymax: 1 ; GradX: 10 ; GradY: 0.1

Observations :

Critère de décision

Pour une étude statistique sur certains caractères connus de la population (âge, sexe, taille, etc), on considère qu'un échantillon est **représentatif**, si

Remarques :

- Il s'agit d'un intervalle dans lequel se situent 95% des échantillons établis dans ces conditions.
  Cet intervalle dépend évidemment de la taille n de l'échantillon et de la probabilité p d'apparition du caractère dans la population.
- Dans la pratique, on connait n et on peut faire en sorte que notre échantillon soit établi au hasard.
  Par contre, pour p, à moins d'avoir regarder la population entière (et dans ce cas, ce chapitre n'a plus d'intérêt) il s'agit d'une hypothèse.
  Cet intervalle de fluctuation servira donc surtout à vérifier la crédibilité d'une hypothèse sur p.
- Ainsi, si la fréquence observée n'appartient pas à l'intervalle de fluctuation au seuil de 95%, on considérera que l'échantillon n'est pas représentatif, ie non établi au hasard, ou encore que l'hypothèse de départ sur p est **mauvaise** et on la rejettera.
  On a un risque de se tromper dans 5% des cas, puisque 5% des fréquences établies dans ces conditions ne sont pas dans cet intervalle. On a donc refusé 5% d'échantillons « en trop ».
  Par contre, si f est dans l'intervalle de fluctuation au seuil de 95%, on considérera l'échantillon comme représentatif, ie établi au hasard, ou encore que notre hypothèse sur p est **crédible**, sans connaître le risque d'erreur, ie le nombre d'échantillons acceptés « en trop ».
  Dans tous les cas, on est sûr de rien ! Donc on évitera le vocabulaire « vrai » ou « faux ».
- Tout est une question d'équilibre :
  Si l'on veut diminuer l'erreur de rejet, par exemple en prenant un intervalle de fluctuation au seuil de 100%, on ne rejettera donc aucun échantillon « en trop », par contre, on les acceptera tous, donc évidemment, beaucoup trop. Toute hypothèse sur p semblera crédible, ce qui n'a aucun intérêt.
  Ainsi, dans la pratique, on utilise surtout les seuils de 95% et de 99%.

En seconde

Pour un caractère donné, on note p sa probabilité d'apparition et f sa fréquence d'apparition dans un échantillon.
La fréquence f se situe dans au moins 95% des cas dans l'intervalle

$$\left[\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots , \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\right]$$

- **Avantages** :

- **Inconvénients** :

Sans parler du fait que cet intervalle n'a jamais été justifié par votre enseignant ...

Applications :

En Première

On appelle X la variable aléatoire qui compte le nombre de personnes d'un échantillon de taille n, possèdant un certain caratère, de probabilité d'apparition p. On a donc $X \hookrightarrow B(n,p)$

On note $f = \dfrac{X}{n}$ la fréquence d'apparition de caractère.

La fréquence f se situe dans au moins 95% des cas dans l'intervalle $\left[\ldots\ldots\ldots, \ldots\ldots\ldots\right]$, où :

– a désigne le plus petit entier k à partir duquel $P(X \leq k)$ dépasse ....
– b désigne le plus petit entier k à partir duquel $P(X \leq k)$ dépasse ....

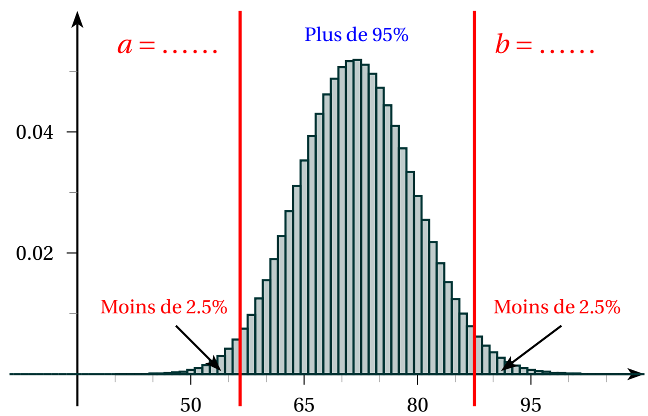
- **Avantages** :

  Sans parler du fait que cet intervalle est justifié par vos connaissances sur la loi binomiale ...
- **Inconvénients** :

Applications : $X \hookrightarrow$

k	$P(X \leq k)$
...	...
54	0.00953
55	0.01375
56	0.01944
57	0.02699
58	0.03678
59	0.04924
...	...

k	$P(X \leq k)$
...	...
83	0.94587
84	0.94587
85	0.95825
86	0.96821
87	0.97609
88	0.98225
...	...



Question 2 : L'étude de Norbert montre que dans cet échantillon, 29% des éléphants sont daltoniens. Estimer la proportion d'éléphants daltoniens dans la population totale.

En Seconde

Pour un caractère donné, on note f sa fréquence d'apparition dans un échantillon de taille n.
Sa probabilité p d'apparition dans la population totale se situe dans au moins 95% des cas dans l'intervalle

$$\left[\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\right]$$

- **Avantages** :
- **Inconvénients** :

  Sans parler du fait que cet intervalle n'a jamais été justifié par votre enseignant ...

Applications :

Exemple :

On admet que dans la population d'enfants de 11 à 14 ans d'un département français le pourcentage d'enfants ayant déjà eu une crise d'asthme dans leur vie est de 13%.

Un médecin d'une ville de ce département est surpris du nombre important d'enfants le consultant ayant des crises d'asthme et en informe les services sanitaires. Ceux–ci décident d'entreprendre une étude et d'évaluer la proportion d'enfants de 11 à 14 ans ayant déjà eu des crises d'asthme.

Ils sélectionnent de manière aléatoire 100 jeunes de 11 à 14 ans de la ville.

La règle de décision prise est la suivante : si la proportion observée est supérieure à la borne supérieure de l'intervalle de fluctuation asymptotique au seuil de 95% alors une investigation plus complète sera mise en place afin de rechercher les facteurs de risque pouvant expliquer cette proportion élevée.

1. Déterminer l'intervalle de fluctuation asymptotique au seuil de 95% de la proportion de jeunes de 11 à 14 ans ayant eu une crise d'asthme dans un échantillon de taille 100.
2. L'étude réalisée auprès des 100 personnes a dénombré 19 jeunes ayant déjà eu des crises d'asthme. Que pouvez-vous conclure ?
3. Le médecin n'est pas convaincu par cette conclusion et déclare que le nombre de personnes interrogées était insuffisant pour mettre en évidence qu'il y avait plus de jeunes ayant eu des crises d'asthme que dans le reste du département.
   Combien faudrait-il prendre de sujets pour qu'une proportion observée de 19% soit en dehors de l'intervalle de fluctuation asymptotique ?

Résumé

	Intervalle de **fluctuation** Au seuil 0.95 (p connue)	Intervalle de **confiance** Au seuil 0.95 (p inconnue)
SECONDE	$n \geq 25$, $0.2 \leq p \leq 0.8$ $\mathrm{I} = \left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}}\right]$	Sensibilisation
PREMIERE	Avec la loi binomiale $\mathrm{I} = \left[\frac{a}{n}; \frac{b}{n}\right]$	
TERMINALE	$n \geq 30$, $np \leq 5$ et $n(1-p) \leq 5$ Asymptotique $\mathrm{I}_n = \left[p - 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$	$\mathrm{J} = \left[f - \frac{1}{\sqrt{n}}, f + \frac{1}{\sqrt{n}}\right]$

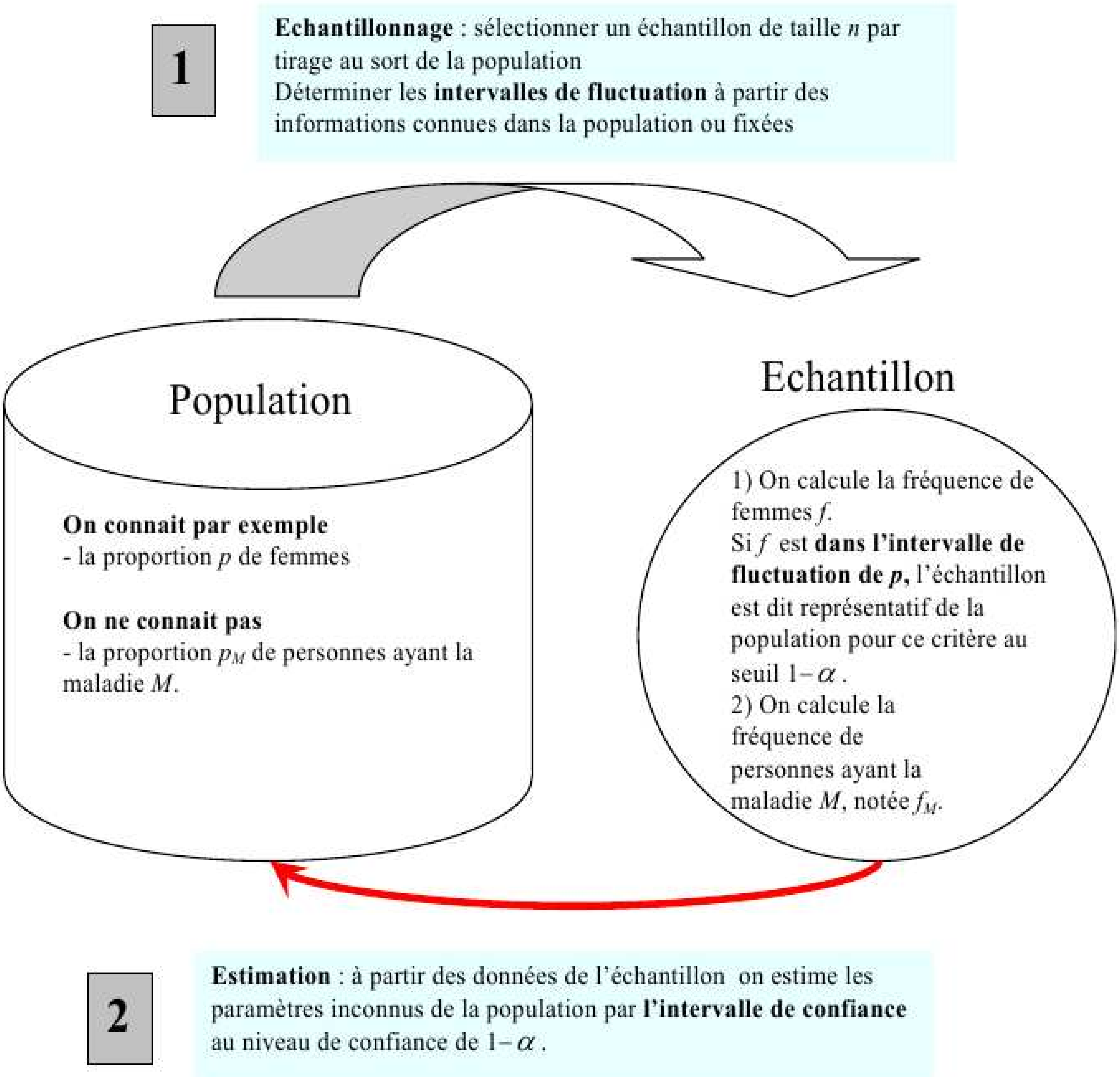
Exemple :

Une entreprise désire contrôler la qualité de sa production.

Un échantillon de 370 produits a été testé. Il ressort de cette étude que 26 produits sur 370 ne sont pas conformes au cahier des charges.

On considère que la production de l'entreprise est suffisamment importante pour qu'on puisse assimiler le tirage de l'échantillon à un tirage avec remise.

1. A partir de cette étude, estimer la probabilité qu'un produit ne soit pas conforme au cahier des charges à l'aide d'un intervalle de confiance au niveau asymptotique 95%.
2. Combien faudrait-il prélever de produit pour avoir un intervalle de confiance au niveau 95% de longueur inférieure à 0.05 ?


1
Echantillonnage : sélectionner un échantillon de taille n par tirage au sort de la population
Déterminer les **intervalles de fluctuation** à partir des informations connues dans la population ou fixées
Population
On connait par exemple
- la proportion p de femmes
On ne connait pas
- la proportion p_M de personnes ayant la maladie M.
Echantillon
1) On calcule la fréquence de femmes f.
Si f est **dans l'intervalle de fluctuation de p,** l'échantillon est dit représentatif de la population pour ce critère au seuil $1-\alpha$.
2) On calcule la fréquence de personnes ayant la maladie M, notée f_M.
2
Estimation : à partir des données de l'échantillon on estime les paramètres inconnus de la population par **l'intervalle de confiance** au niveau de confiance de $1-\alpha$.