Chapitre 7 - Échantillonnage

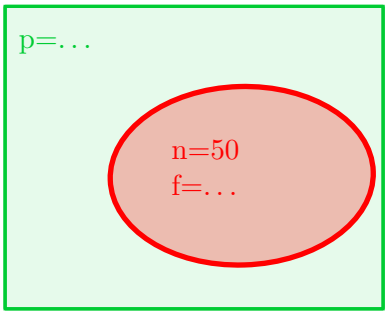
I Échantillonnage

TD : Observer des échantillons

Dans une usine fabriquant des automobiles, on contrôle les défauts de peinture de type "grains ponctuels".

On considère que ce défaut, presque invisible, touche 20% de la production.

1. Quelle est la **population** étudiée ici ? Quel caractère de la population observe-t-on ? Quelle est, sous forme décimale, sa fréquence p ?
2. Un ingénieur-contrôle procède à 50 tirages au haard d'une voiture dans la production et note, pour chacune, si le défaut est présent (1) ou non (0). La liste obtenue, comprenant 50 nombres 0 ou 1 est un **échantillon de taille 50**.
   (a) Sur cet échantillon, le 1 apparaît quatre fois. Calculer la fréquence f du défaut sur cet échantillon. Compléter le schéma ci-dessous pour résumer cette prise d'échantillon.



   (b) On donne les résultats obtenus sur 8 autres échantillons de taille 50 de cette même production.

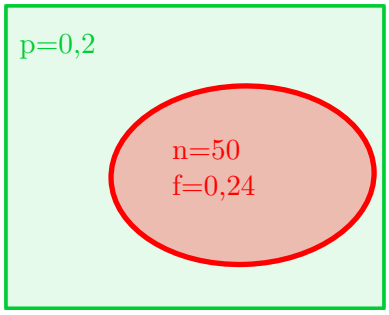
Échantillon n°	1	2	3	4	5	6	7	8
Nombre de défauts	11	9	16	9	11	11	10	4
Fréquence du défaut								

   Compléter ce tableau. Retrouve-t-on à chaque fois la fréquence p connue dans la population ? Comment s'appelle ce phénomène ?
3. **L'année suivante,** un nouvel échantillon de taille 50 révèle 24% de défauts. Doit-on nécessairement en conclure que la qualité a baissé ?

Définition 1

*Un **échantillon (aléatoire) de taille** n est la liste des résultats obtenus par n répétitions indépendantes d'une même expérience aléatoire.*

Exemple : Si un échantillon compte 12 fois le 1 (et donc 34 fois le 0) la fréquence f du caractère "défaut" sur cet échantillon est égale à $\frac{12}{50} = 0,24$. Le schéma ci-dessous en est un résumé.



Si l'on réalise plusieurs échantillons de taille 50, la fréquence f de "défaut" varie d'un échantillon à l'autre et n'est pas égale en général à la fréquence $p = 0,2$ de "défaut" dans la population.
On dit que f **fluctue** autour de p ; il s'agit de la **fluctuation d'échantillonnage**

II Intervalle de fluctuation

TP : Chroniques d'un village chinois
Dans un village des montagnes chinoises en 2000, il est né 25 enfants dont 20 garçons. On se pose la question de savoir si le hasard seul peut "raisonnablement" expliquer cette observation statistique. Pour cela, on étudie la fréquence des garçons sur des échantillons de taille 25 que l'on simule en supposant l'équiprobabilité des sexes à la naissance.

1. **Avec une pièce supposée bien équilibrée**
   (a) Lancer 25 fois une pièce de monnaie et calculer la fréquence de "pile".
   (b) En regroupant dans un tableau les fréquences de "pile" obtenues par les élèves de la classe, calculer le pourcentage des échantillons de 25 lancers ayant donné une fréquence de "pile" comprise entre $0,3$ et $0,7$.
2. **Avec un tableur** Ouvrir le fichier echantillonnage1.ods. Recopier la colonne B vers la droite de façon à simuler 100 échantillons de 25 naissances et à disposer des fréquences de "garçon" associées. Sélectionner la dernière ligne et créer le nuage de points associé à ces fréquences. Déterminer le pourcentage des fréquences f telles que $0,3 \leqslant f \leqslant 0,7$.
3. **Intervalles de fluctuation de f**
   (a) Pour $n = 25$, préciser les bornes de l'intervalle $I = \left[0,5 - \frac{1}{\sqrt{n}} \ ; \ 0,5 + \frac{1}{\sqrt{n}}\right]$.
   (b) Appuyer sur F9 un grand nombre de fois. Estimer le pourcentage des échantillons de taille 25 donnant une fréquence de "garçon" dans cet intervalle.
4. **Prise de position**
   (a) Calculer la fréquence des garçons nés dans le village chinois en 2000.
   (b) Que peut-on penser à propos des naissances dans ce village chinois ?

TD : Parité à l'embauche ?

Dans un secteur d'activité où il y a autant de femmes que d'hommes, une petite entreprise A de 35 personnes emploie 40% de femmes, alors qu'une importante société B compte 46% de femmes parmi ses 3 500 salariés.

1. A votre avis, laquelle de ces deux entreprises respecte le moins bien la parité hommes/femmes ?
2. On a simulé 100 échantillons de taille 35 (figure 1) et 100 échantillons de taille 3 500 (figure 2) de lancers d'une pièce supposée équilibrée. Les points correspondent à la fréquence de pile sur chaque échantillon.

figure 1

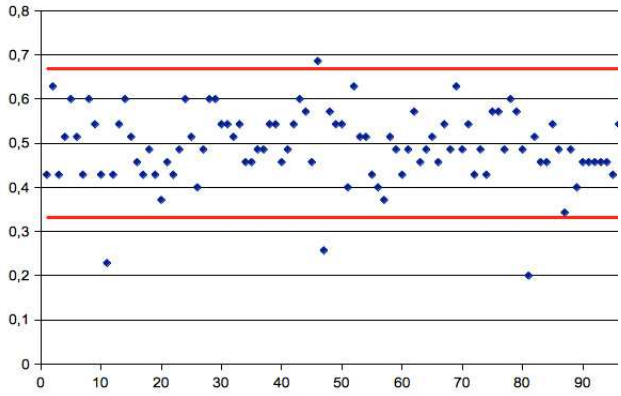
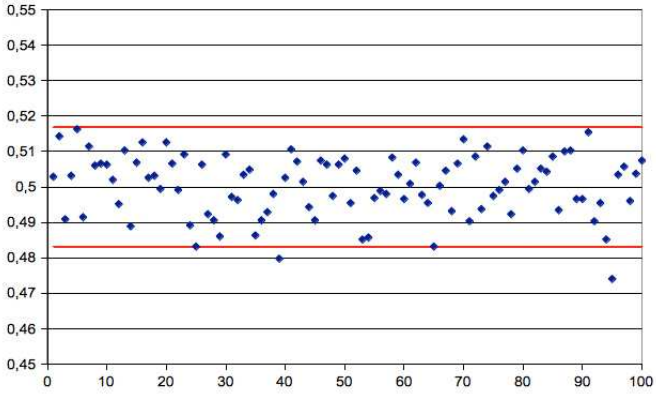


figure 2



   (a) Calculer les bornes de l'intervalle $I_1 = \left[0,5 - \frac{1}{\sqrt{35}} \ ; \ 0,5 + \frac{1}{\sqrt{35}}\right]$ et le visualiser sur la figure 1. Combien de fréquences se trouvent en dehors de cet intervalle ? Peut-on dire, ici, que pour 95% au moins de ces échantillons de taille 35, la fréquence de "pile" appartient à cet intervalle ?

   (b) Reprendre la question précédente avec $I_2 = \left[0,5 - \frac{1}{\sqrt{3\ 500}} \ ; \ 0,5 + \frac{1}{\sqrt{3\ 500}}\right]$, "figue 2" et "taille 3 500".

3. Que peut-on dire maintenant des proportions de femmes dans les sociétés A et B ? Êtes-vous toujours en accord avec la réponse donnée à la question 1 ?

Propriété 1 (Admise et non exigible)

Lorsqu'on prélève un échantillon de taille n dans une population où la fréquence d'un caractère est p, alors sous certaines conditions, la probabilité que cet échantillon fournisse une fréquence appartenant à l'intervalle $I = \left[p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}}\right]$ est au moins égale à $0,95$.

Définition 2

I est appelé ***intervalle de fluctuation de la fréquence f au seuil de*** 95%.

TP : Un cas de discrimination ?

En 1977 a été portée devant la Cour suprême des États-Unis une affaire où le gouvernement fédéral suspectait l'école indépendante d'Hazelwood, située dans la banlieue de Saint-Louis, de discrimination à l'embauche à l'égard des professeurs afro-américains. Pour prendre sa décision, la Cour disposait des statistiques suivantes (H. Zeisel, D. Kaye, *Prove it with Figures*) :

– sur la période 1972-74, cette école a employé 405 professeurs dont 15 Afro-américains ;
– le pourcentage d'afro-américains sur le marché du travail correspondant était de $15,4\%$, en incluant la ville de Saint-Louis, et de $5,7\%$ en l'excluant.

A. Intervalle de fluctuation sous l'hypothèse $p = 0,154$

1. Appliquer les instructions tableur ci-dessous :
   – Entrer en cellule B1 la valeur 0,154
   – Entrer en A3 la formule =ENT(ALEA()+B1)
   – Recopier vers le bas jusqu'en A407
   – En cellule A408, entrer =SOMME(A3 :A407)/405
   – Sélectionner les cellules de A3 à A408 puis recopier vers la droite jusqu'en colonne CV

   Préciser :
   (a) à quoi correspondent les valeurs 0 et 1 ;
   (b) ce que représente la liste de ces 405 "0 ou 1" ;
   (c) ce que représente le contenu de la cellule A408 ;
   (d) quel travail statistique a été réalisé ici.
2. Représenter les fréquences simulées par un nuage de points.
   Observer l'effet de la touche F9.
3. Poursuivre avec les instructions ci-dessous :
   – Entrer en E1 la formule =B1-1/RACINE(405) et en G1 la formule =B1+1/RACINE(405)
   – Entrer en I1 la formule =NB.SI(A408 :CV408 ;">="&E1)-NB.SI(A408 :CV408 ;">"&G1)

   (a) de quel intervalle obtient-on les bornes en E1 et en G1 ?
   (b) préciser quel test on effectue sur les fréquences et quel est le nombre obtenu en cellule I1. Observer l'effet de la touche F9.
4. Calculer en B2 la fréquence f des professeurs afro-américains employés à l'école d'Hazelwood. Comparer aux fréquences des échantillons simulés.

B. Intervalle de fluctuation sous l'hypothèse $p = 0,057$

Dans le quartier où se situe l'école, il n'y a plus que $5,7\%$ de professeurs afro-américains potentiels.

1. Modifier la valeur en cellule B1 et donner, dans ce cas, l'intervalle de fluctuation au seuil 0,95.
2. Observer si la fréquence f appartient à cet intervalle et expliquer la décision des juges de la Cour suprême, en faveur de l'école d'Hazelwood, par 8 votes contre 1.

Exercice : Qui croire ?

Au 1^{er} tour de l'élection présidentielle de 2002, les résultats sont : Jacques Chirac : $19,88\%$, Lionel Jospin : $16,18\%$ et Jean-Marie Le Pen : $16,86\%$.

Le lendemain, le 22 avril 2002, on pouvait dans le journal Libération : "L'éviction de Jospin laisse Le Pen face à Chirac ! Un séisme. Le duel entre Jacques Chirac et Lionel Jospin n'aura pas lieu."

Le dernier sondage effectué sur environ 1000 personnes, publié par BVA le 19/04/2002 donnait comme prévisions : Jacques Chirac : 19%, Lionel Jospin : 18% et Jean-Marie Le Pen : 14% !

Unanimement les médias mettent en cause les sondages. A contre-courant, un statisticien s'exprime dans le journal Le Monde : "Pour les rares scientifiques qui savent comment sont produites les estimations, il est clair que l'écart des intentions de vote entre les candidats Le Pen et Jospin rendait tout à fait plausible le scénario qui s'est réalisé."

Alors qui croire ?

1. Déterminer, pour chaque candidat, l'intervalle de confiance $I = \left[p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}}\right]$ au seuil de 95%. Interpréter ces intervalles et les représenter l'un en-dessous de l'autre.
2. Entre les deux points de vue contradictoires exprimés, lequel adoptez-vous ? Justifier.