Chapitre 3 ÉCHANTILLONNAGE

3.1 Introduction

La démarche d'échantillonnage est une démarche statistique classique de type déductif c'est à dire qui va du "général au particulier" : on connaît la population, on s'intéresse à l'échantillon. Prenons trois exemples.

On connaît les professions d'une population cible dans laquelle est prélevé un échantillon. Est-ce que cet échantillon peut être considéré comme représentatif de la population selon la variable profession ?

On s'intéresse au contrôle de la qualité de fabrication de tablettes de chocolat. Est-ce qu'on peut considérer comme constant le poids moyen garanti d'une tablette ? Pour cela, on prélève régulièrement un échantillon de n tablettes dont l'étude statistique permettra de répondre à la question.

Dans la fabrication d'aliment pour poulets conditionné en sacs de 10 kilos, on indique sur les sacs la composition de l'aliment (proportions des composants). Des échantillons sont prélevés sur les lieux de vente pour contrôler le respect de ces indications.

3.2 Concept de base des distributions d'échantillonnage

3.2.1 Distribution d'échantillonnage des moyennes et des variances

Soit une population constituées de N individus, on s'intéresse a une variable aléatoire X, on prélevé p échantillon ξ de taille n, on dispose alors de plusieurs série statistique de moyenne et variance calculable

Nous rappelons les caractéristiques de la population :

- taille N (finie ou infinie)

- X = variable aléatoire quelconque

- $E(X)=m_o$

- $Var(X)=\sigma^2$

Échantillons	Valeurs observées	Moyennes observées	Variances observées(empiriques)
$\mathscr{E}_1$	$x_{11}, x_{12}, \ldots, x_{1n}$	$\overline{x_1}$	s'^2_1
$\mathscr{E}_2$	$x_{21}, x_{22}, \ldots, x_{2n}$	$\overline{x_2}$	s'^2_2
...	...	...	...
$\mathscr{E}_p$	$x_{p1}, x_{p2}, \ldots, x_{pn}$	$\overline{x_p}$	s'^2_p
Variables aléatoires	$X_1 \quad X_2 \quad \ldots \quad X_n$	$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$	$S'^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$

Les distributions des variables aléatoires $\overline{X}$ et S^2 sont appelées d'échantillonnage des moyennes et des variances.

3.3 Distribution d'échantillonnage d'une variance dans le cas d'une population normale

3.3.1 Présentation des données et position du problème

Dans une chocolaterie, on étudie la fiabilité d'un procédé de fabrication de tablettes de chocolat de 100 g et l'on veut, bien entendu, s'assurer la maîtrise de la variabilité de ce poids.

On note X, la variable aléatoire "poids d'une tablette fabriquée". Lorsque toute la chaîne fonctionne correctement, l'écart-type est égal à 5 g. Dans ce type d'application, on considère la variable aléatoire X distribuée selon une loi normale.

Afin de contrôler la variabilité, on prélève périodiquement un échantillon de 10 tablettes et on en calcule la variance observée s^2

3.3.2 Les Questions

• Déterminer l'intervalle $[S_a'^2 , S_b'^2]$ qui a une sécurité de 95% de contenir la variance S'^2 observée dans un tel échantillon. Cet intervalle est dit "intervalle de probabilité" ou "intervalle de pari"(noté IP). Le risque 5% est notéα.

• Étendre ces calculs aux cas suivants :

- réduction du risque α aux valeurs 3%, 1% et 3‰

- échantillons de tailles n = 20 puis 30 tablettes

- étude du cas d'un écart-type σ = 3 g correspondant à l'acquisition d'une machine plus performante.

3.3.3 Notations et modèle

• Population : c'est l'ensemble de tablettes de 100 g fabriquées par la société.

- X est la variable aléatoire, poids d'une tablette

- E(X) = m_o est le poids moyen d'une tablette

- Var X = σ^2

$-X \to N(m, \sigma)$

• echantillon

- La taille est n (ici n= 10)

- $X_1, X_2, \ldots, X_n$ sont des variables aléatoires indépendantes

- $X_1 \to N(m, \sigma) \quad \forall i \in \{1,2, \ldots., n\}$

$S'^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$ est la variable aléatoire variance dans un échantillon de taille n

3.3.4 Les Resultats :

Intervalle de probabilité ou de pari de la variance de l'échantillon au niveau 1-α est

$$\frac{\sigma^2}{n}\ \chi^2{}_{(n-1),\frac{\alpha}{2}} \leq\ S'^2\ \leq \frac{\sigma^2}{n}\ \chi^2{}_{(n-1),1-\frac{\alpha}{2}}$$

On utilise la table de la loi de khi 2 , avec (n-1) degré de liberté , et la probabilité

$$\chi^2{}_{9,0.025} = 2{,}70$$

$$\chi^2{}_{9,0.975} = 19{,}0227$$

Donc :

$$\frac{25}{10}\ 2{,}70 \leq\ S'^2\ \leq \frac{25}{10}\ 19{,}0227$$

$$\mathbf{6,75} \leq\ \boldsymbol{S'^2}\ \leq \mathbf{47,56}$$

On en déduit que lorsque la chaîne de production fonctionne correctement, la variance observée dans un échantillon de 10 tablettes à 95% de chances d'être comprise entre 6,75 et 47,56.

En faisant varier les calculs suivant les paramètres suivant :

- réduction du risque α aux valeurs 3%, 1% et 3‰
- échantillons de tailles n = 20 puis 30 tablettes
- étude du cas d'un écart-type σ = 3 g correspondant à l'acquisition d'une machine plus performante.

Nous obtenons le tableau suivant :

α	σ^2	n	$\chi^2{}_a$	$\chi^2{}_b$	s'^2_a	s'^2_b
5,0%	25	10	2,70	19,02	6,75	47,56
3,0%	25	10	2,33	20,51	5,84	51,28
1,0%	25	10	1,73	23,59	4,34	58,97
0,3%	25	10	1,27	26,82	3,19	67,04
5,0%	25	20	8,91	32,85	11,13	41,07
3,0%	25	20	8,16	34,74	10,20	43,43
1,0%	25	20	6,84	38,58	8,55	48,23
0,3%	25	20	5,73	42,53	7,16	53,17
5,0%	25	30	16,05	45,72	13,37	38,10
3,0%	25	30	15,00	47,91	12,50	39,93
1,0%	25	30	13,12	52,34	10,93	43,61
0,3%	25	30	11,47	56,84	9,56	47,37

α	σ^2	n	$\chi^2{}_a$	$\chi^2{}_b$	s'^2_a	s'^2_b
5,0%	9	10	2,70	19,02	2,43	17,12
3,0%	9	10	2,33	20,51	2,10	18,46
1,0%	9	10	1,73	23,59	1,56	21,23
0,3%	9	10	1,27	26,82	1,15	24,14
5,0%	9	20	8,91	32,85	4,01	14,78
3,0%	9	20	8,16	34,74	3,67	15,63
1,0%	9	20	6,84	38,58	3,08	17,36
0,3%	9	20	5,73	42,53	2,58	19,14
5,0%	9	30	16,05	45,72	4,81	13,72
3,0%	9	30	15,00	47,91	4,50	14,37
1,0%	9	30	13,12	52,34	3,94	15,70
0,3%	9	30	11,47	56,84	3,44	17,05

Tableau : variation de l'intervalle de probabilité de la variance observée selon le risque , la taille d l'échantillon , la variance de la population.

Nous avons ci-dessous un exemple de table pour la loi khi2

	α												
dl	0.5%	1%	2.5%	5%	10%	25%	50%	75%	90%	95%	97.5%	99%	99.5%
1	0.00	0.02	0.00	0.00	0.02	0.10	0.46	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.21	0.05	0.10	0.21	0.57	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.58	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.35	12.84
4	0.21	1.06	0.48	0.71	1.06	1.92	3.36	5.38	7.78	9.49	11.14	13.28	14.86
5	0.41	1.61	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	2.20	1.24	1.64	2.20	3.46	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	2.83	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	3.49	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.54	20.09	21.95
9	1.73	4.17	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	4.87	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	5.58	3.82	4.58	5.58	7.58	10.34	13.70	17.27	19.68	21.92	24.73	26.76
12	3.07	6.30	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30
13	3.56	7.04	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.08	7.79	5.63	6.57	7.79	10.16	13.34	17.12	21.06	23.68	26.12	29.14	31.32

3.3.5 Commentaire des résultats

Bien entendu, on retrouve des résultats conformes à la formule mathématique.

- Pour une variance σ^2 et un risque α donnés, l'intervalle de probabilité IP est plus resserré si l'on augmente la taille de l'échantillon

- pour une variance σ^2 et une taille d'échantillon n données, l'intervalle de probabilité IP augmente lorsque le risque diminue

- pour une taille et un risque α donnés, l'intervalle de probabilité IP diminue si l'on diminue la variance.

En examinant ces résultats, on peut par exemple porter son attention sur le risque 3 ‰ fréquemment adopté dans l'industrie, sur un échantillon de taille 10 et une variance de 25.

L'intervalle trouvé pour la variance de l'échantillon [3,19 ; 67,04] est "vaste". Il se resserre sensiblement avec un échantillon de taille 20 : [7,16 ; 53,17]. Enfin, on note une bonne précision, si la variance liée à l'ensemble du processus de fabrication peut être ramenée à 9 avec un échantillon de taille 30 puisque alors, la fourchette se réduit à [3,44,17,05].

Lorsque l'échantillonnage ne détruit pas l'objet, il est souvent intéressant de prélever des échantillons de taille plus importante.

3.4 DISTRIBUTION D'ÉCHANTILLONNAGE D'UNE MOYENNE : Population normale de moyenne et variance connues

3.5 Exemple : variabilité du poids de tablettes de chocolat

3.5.1 Présentation des données et position du problème

On se place dans le même environnement concret que dans l'étude précédente (échantillonnage d'une variance). Dans la fabrique de chocolats, le service qualité s'intéresse à la qualité de remplissage des tablettes. Lorsque le fonctionnement de la chaîne est correct, le poids d'une tablette est une variable aléatoire X normale, de moyenne m = 100 g et d'écart type σ = 5g.

Le contrôle est réalisé en prélevant périodiquement sur la chaîne un échantillon de n = 10 tablettes. Concrètement, on calcule le poids moyen x observé dans un tel échantillon et l'on examine s'il ne s'écarte "pas trop" du poids moyen théorique de 100 g, ou encore, s'il appartient à une fourchette de poids "jugée" convenable ou enfin, dans certains cas, s'il reste supérieur à un poids minimum garanti.

3.5.2 Question 1

a) A quel intervalle [$\overline{X_a}, \overline{X_b}$] dit "intervalle de probabilité" ou "intervalle de pari" doit appartenir le poids moyen d'une tablette dans un tel échantillon avec un niveau de sécurité de 1-α = 0,95 (α = 5% est le risque). Noter que cette question équivaut à rechercher l'écart Δ tel que la moyenne d'échantillon appartienne à l'intervalle [1OO-Δ ; 100+Δ] avec une probabilité 1-α

b) Quel poids moyen minimum G peut-on garantir au risque α ?

3.5.3 Question 2

II est intéressant d'étudier l'évolution de la précision Δ et par suite celle de l'IP en faisant varier le risque, la taille de l'échantillon et même la variance σ^2

Étendre les calculs réalisés à la question 1 aux cas suivants :

- réduction du risque a aux valeurs 3%, 1% et 3 ‰ (remarque : dans l'industrie, les risques sont souvent très petits car on ne souhaite retoucher au processus que lorsque c'est vraiment nécessaire)

- échantillon de tailles n = 20 et 30

- écart-type σ = 3, correspondant par exemple à l'acquisition d'une nouvelle machine de variabilité réduite.

3.5.4 Notations et modèle

• Population : c'est l'ensemble de tablettes de 100 g fabriquées par la société.

- X est la variable aléatoire, poids d'une tablette

- E(X) est le poids moyen d'une tablette

- Var X = σ^2

$-X \rightarrow N(m, \sigma)$

• echantillon

- La taille est n (ici n= 10)

- $X_1, X_2, \ldots, X_n$ sont des variables aléatoires indépendantes

- $X_1 \rightarrow N(m, \sigma) \quad \forall i \in \{1,2, \ldots., n\}$

3.5.5 Le calcul statistique

La distribution de la moyenne d'échantillonnage est :

$E(\overline{X}) = m \quad , Var(\overline{X}) = \frac{\sigma^2}{n}$

$\overline{X}$ suit la loi de probabilité : $\overline{X} \rightarrow N(m, \frac{\sigma}{\sqrt{n}})$

3.5.5.1 Question 1a :

On cherche l'intervalle [$\overline{X_a}, \overline{X_b}$] tel que $P(\overline{X_a} \leq \overline{X} \leq \overline{X_b}) = 1 - \alpha$

Autrement dit, on cherche Δ tel que P(m - Δ $\leq \overline{X} \leq$ m + Δ) = $1 - \alpha$ (le risque est réparti sur les deux queues de la distribution).

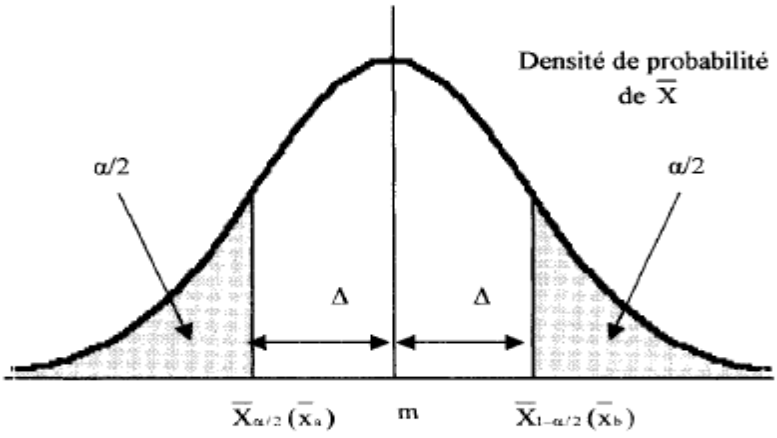

Figure : distribution de la moyenne d'échantillonnage

$\frac{\alpha}{2} = 0.025$ Donc on recherche des deux de la table de la loi normale centrée les valeurs U correspondante a 0.025 pour la valeur minimale de l'intervalle et 0.975 pour la valeur maximale

Pour 0.975 la table donne U=1.96 comme il y a symétrie on $U_a = -1.96 \; et \; U_b = 1.96$

$$\frac{X_a - m}{\frac{\sigma}{\sqrt{n}}} = U_a \quad donc \; X_a = m + \left(\mathrm{U}.\frac{\sigma}{\sqrt{n}}\right) = m - \Delta$$

$$\frac{X_b - m}{\frac{\sigma}{\sqrt{n}}} = U_b \quad donc \; X_b = m + \left(U_b.\frac{\sigma}{\sqrt{n}}\right) = m + \Delta$$

$$X_a = m + \left(U_a.\frac{\sigma}{\sqrt{n}}\right) = m - \Delta = 60 + \left(\frac{-1{,}96.0{,}5}{\sqrt{10}}\right) = 60 - 0{,}31 = 59{,}7 \; g$$

$$X_b = m + \left(U_b.\frac{\sigma}{\sqrt{n}}\right) = m + \Delta = 60 + \left(\frac{1{,}96.0{,}5}{\sqrt{10}}\right) = 60 + 0{,}31 = 60{,}31 \; g$$

$$\mathbf{59{,}7 \leq \overline{X} \leq 60{,}31}$$

3.5.5.2 Interprétation

Lorsque le processus de fabrication fonctionne correctement, en prélevant un échantillon de 10 tablettes, on peut "parier" que le poids moyen d'une tablette dans cet échantillon a 95% de chances d'appartenir à l'intervalle [96,90; 103,1] ou encore que ce poids moyen est de 100g avec une erreur maximale de 3,1 g au risque de 5%.

3.5.5.3 Question I- b

On cherche G tel que $P(\overline{X} \geq G) = 1 - \alpha$

G est le fractile d'ordre α de la loi de probabilité de $\overline{X}$, c'est à dire de la loi $N(m, \frac{\sigma}{\sqrt{n}})$

$\alpha = 0.05$ donc on recherche des deux de la table de la loi normale centrée la valeur U_G inferieur correspondante a 0.05

Pour 0.05 la table donne U_G = -1.645

$$\frac{X_G - m}{\frac{\sigma}{\sqrt{n}}} = \mathrm{U}_G \quad donc \; X_G = m + \left(\mathrm{U}_G.\frac{\sigma}{\sqrt{n}}\right) = 100 - \left(\frac{1{,}645.5}{\sqrt{10}}\right) = 100 - 2{,}6 = 97{,}4 \; gr$$

En faisant varier les calculs suivant les paramètres suivant :

- réduction du risque α aux valeurs 3%, 1% et 3‰

- échantillons de tailles n = 20 puis 30 tablettes

- étude du cas d'un écart-type σ = 3 g correspondant à l'acquisition d'une machine plus performante.

α	Niveau sécurité (1-α)	σ	n	Δ fonction IC	$\overline{x_a}$	$\overline{x_b}$	G (poids moyen minimum garanti)
5,00%	95,00%	5	10	3,10	96,90	103,10	97,40
3,00%	97,00%	5	10	3,43	96,57	103,43	97,03
1,00%	99,00%	5	10	4,07	95,93	104,07	96,32
0,30%	99,70%	5	10	4,69	95,31	104,69	95,66
5,00%	95,00%	5	20	2,19	97,81	102,19	98,16
3,00%	97,00%	5	20	2,43	97,57	102,43	97,90
1,00%	99,00%	5	20	2,88	97,12	102,88	97,40
0,30%	99,70%	5	20	3,32	96,68	103,32	96,93
5,00%	95,00%	5	30	1,79	98,21	101,79	98,50
3,00%	97,00%	5	30	1,98	98,02	101,98	98,28
1,00%	99,00%	5	30	2,35	97,65	102,35	97,88
0,30%	99,70%	5	30	2,71	97,29	102,71	97,49
5,00%	95,00%	3	10	1,86	98,14	101,86	98,44
3,00%	97,00%	3	10	2,06	97,94	102,06	98,22
1,00%	99,00%	3	10	2,44	97,56	102,44	97,79
0,30%	99,70%	3	10	2,82	97,18	102,82	97,39
5,00%	95,00%	3	20	1,31	98,69	101,31	98,90
3,00%	97,00%	3	20	1,46	98,54	101,46	98,74
1,00%	99,00%	3	20	1,73	98,27	101,73	98,44
0,30%	99,70%	3	20	1,99	98,01	101,99	98,16
5,00%	95,00%	3	30	1,07	98,93	101,07	99,10
3,00%	97,00%	3	30	1,19	98,81	101,19	98,97
1,00%	99,00%	3	30	1,41	98,59	101,41	98,73
0,30%	99,70%	3	30	1,63	98,37	101,63	98,49

Figure : détermination de l'intervalle de probabilité du poids moyen et du poids moyen minimum garanti au risque α.Evolution de ces resultats en fonction de α , σ et n .

3.6 Interprétation

Pour une même taille d'échantillon, Δ (erreur absolue) augmente lorsque le risque diminue. Par exemple, pour un échantillon de 10 tablettes au risque de 3%, il conviendra de réviser la chaîne de production dès que le poids moyen d'un tel échantillon s'écarte de plus de 3,43 g de la référence 100 g. Si le risque accepté est 10 fois plus petit, soit 3‰ , on n'effectuera ce contrôle que si l'écart à la référence est beaucoup plus net (4,69 g).

Pour un risque donné, augmenter la taille de l'échantillon augmente la précision et donc diminue A. Ainsi, au risque 3‰ évoqué ci-dessus, avec un échantillon de 20 tablettes, l'écart Δ n'est plus que de 3,32 g contre 4,69 g pour 10 tablettes. Cet écart, révélateur d'une probable avarie de la chaîne de production, n'est plus que de 2,71 g avec un échantillon de 30 tablettes.

Quand l'échantillonnage ne détruit pas l'objet prélevé et n'est pas trop coûteux en temps, il est donc intéressant d'augmenter la taille.

Bien entendu, l'amélioration du fonctionnement de la chaîne visant à diminuer la variabilité va dans le même sens. Avec un écart-type de 3 (au lieu de 5), nous trouvons qu'avec un risque de 3‰ et un échantillon de 30, il suffit de détecter un écart de 1,63 g pour être amené à effectuer une révision de la chaîne. Rappelons que l'écart était de 2,71 avec l'écart-type a = 5.

Résumé de calcul :

Distribution d'échantillonnage d'une variance dans le cas d'une population normale

- Déterminer l'intervalle $[S_a'^2 , S_b'^2]$
- Intervalle de probabilité ou de pari de la variance de l'échantillon au niveau 1-α est :

$$\frac{\sigma^2}{n}\ \chi^2_{(n-1),\frac{\alpha}{2}} \leq S'^2 \leq \frac{\sigma^2}{n}\ \chi^2_{(n-1),1-\frac{\alpha}{2}}$$

On utilise la table de la loi de khi 2 , avec (n-1) degré de liberté , et la probabilité correspondante

DISTRIBUTION D'ÉCHANTILLONNAGE D'UNE MOYENNE : Population normale de moyenne et variance connues

La distribution de la moyenne d'échantillonnage est :

$\mathrm{E}(\overline{X})$=m , $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$

$\overline{X}$ suit la loi de probabilité : $\overline{X} \rightarrow N(m, \frac{\sigma}{\sqrt{n}})$

P(m - $\Delta \leq \overline{X} \leq$ m + Δ) = $1 - \alpha$

$$\frac{X_a - m}{\frac{\sigma}{\sqrt{n}}} = U_a \quad donc\ X_a = m + \left(U_a . \frac{\sigma}{\sqrt{n}}\right) = m - \Delta$$

$$\frac{X_b - m}{\frac{\sigma}{\sqrt{n}}} = U_b \quad donc\ X_b = m + \left(U_b . \frac{\sigma}{\sqrt{n}}\right) = m + \Delta$$