CHAP III NOTION DE VARIABLES ALEATOIRES ET DE LOI DE PROBABILITE

1. Généralité

Soit une épreuve probabiliste engendrant des événements E1, E2  ……….En

Si à chacune des événements, on fait correspondre un nombre, on dit que ce nombre est une variable aléatoire.

X = (x1, x2……………, xn)

Pour définir complètement une variable aléatoire, il faut définir l’ensemble des valeurs possibles qu’elle peut prendre, ainsi que les probabilités correspondantes.

Exemple : Dans un jet de dé, l’expérience donne lieu à six résultats distincts. Si le dé est parfaitement équilibré, chacun des résultats a pour probabilité .

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X | 1 | 2 | 3 | 4 | 5 | 6 |
| pi |  |  |  |  |  |  |

On dit que X est une variable aléatoire et l’ensemble des couples (xi, pi) constitue par définition la loi de probabilité de la variable aléatoire X.

1. Variable aléatoire continue - Fonction de répartition

Au lieu de prendre seulement des variables aléatoires discrètes, une variable aléatoire peut dans certain cas prendre toutes les valeurs d’un intervalle fini ou infini.

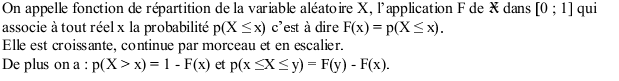
Exemple : Considérons la taille des habitants d’une ville quelconque. La taille est une variable aléatoire qui peut prendre à priori n’importe quelle valeur entre un minimum m et une valeur maximum M.

Une telle variable aléatoire est appelée variable aléatoire continue. Sa loi de probabilité est déterminée si l’on connait pour tout intervalle (a,b) la probabilité que X soit comprise entre a et b, plus précisément :

Pr(a X b)

Cette loi de probabilité est complètement déterminée par la connaissance de la fonction

F(x) = Pr(X x) appelée fonction de répartition de la variable aléatoire X.



En effet pour a b le théorème des probabilités totales permet d’écrire :

P(X b) = P(X a) + P( a X b)

ou

F(b) = F(a) + P( a X b)

ou encore

P( a X b) = F(b) – F(a)

Toute fonction de répartition possède les trois propriétés suivantes :

- F(+) = 1

- F(-) = 0

- F(x) est une fonction croissante de x

2- Valeurs caractéristiques d’une variable aléatoire

1. Variable aléatoire discrète

- Espérance mathématique



La variable aléatoire x – E(x) est appelée variable aléatoire centrée. Son espérance mathématique est nulle.

E( x - E( x)) = 0

- Variance et écart-type

On donne le nom de variance ou moment centré d’ordre 2 et on note :

V(x) = 2  = (xi – m)2 = E(x2) – [E(x)]2



1. Variable aléatoire continu

L’espérance mathématique de X est défini par :

m = E(x) = f(x)dx

La variance de X est par définition :

V(x) = =

La variance peut être exprimée au moyen des moments non centrés d’ordre 1 et d’ordre 2 comme dans le cas de variable discrète.

2 = E(x2) – [E(x)]2

1. Variables aléatoires simultanées

Considérons le jeu simultané de deux dés. Les valeurs indiquées par chacun des deux dés sont deux variables aléatoires X et Y. On dit que ce sont deux variables aléatoires simultanés.

E(x + y) = E(x) + E(y)

V(x + y) = V(x) + V(y)

Exercice d’application :

Neuf chevaux, quatre blancs et cinq noirs pénètrent un à un sur la piste d’un cirque. Ils sont distincts les uns par rapport aux autres par leur équipement.

1. Combien existe -t- il de façons de les arriver sur la piste ?
2. On appelle X la variable aléatoire : nombre de chevaux blancs précédant le premier noir. Donner la loi de probabilité X
3. Calculer l’espérance mathématique de X et sa variance.

Solution :

1. Tout défilé de 9 chevaux correspond à un permutation d’un ensemble de 9unités. Il y a donc 9! façons de faire arriver sur la piste les neuf chevaux.

9 ! = 382 880 façons

1. Soit X la variable aléatoire nombre de chevaux blancs précédant le premier cheval noir. On a X = (0, 1, 2, 3, 4)
2. Pour X = 0, il y a 5 façons de débuter le défilé par un cheval noir. Il restera ensuite 4 chevaux blancs et 4 noirs que l’on peut permuter de toutes les façons possibles.

n0 = 5 x 8!

d’où P(x = 0) = =

1. Pour X = 1, c'est-à-dire un cheval en tête. Il y a 4 façons possibles et un cheval noir en 2ème position peut se produire de 5 façons possibles. Il reste 7 chevaux que l’on peut permuter de toutes façons possibles.

d’où n1  = x x 7!

P(x = 1) = = =

1. Pour X = 2 deux chevaux blancs doivent être en tête , donc façons possibles. Un cheval noir étant en 3ème position de façons possibles, les 6 autres chevaux pouvant être permutés de toutes les façons possibles.

d’où n2  = x x 6!

P(x = 2) = = =

1. Pour X = 3 trois chevaux blancs doivent être en tête , donc façons possibles. Un cheval noir étant en 4ème position de façons possibles, les 5 autres chevaux pouvant être permutés 5 ! façons possibles.

d’où n3  = x x 5!

P(x = 3) = =

1. Pour X = 4 par un raisonnement analogue

d’où n4  = x x 4!

P(x = 4) = =

La loi de probabilité de la variable aléatoire X se résume comme suit :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| k | 0 | 1 | 2 | 3 | 4 | Total |
| P(X = k) |  |  |  |  |  | 1 |

1. Calcul de E(X) et V(X)

E(X) =

= =

E(X) =

V(X) = = -

V(X) =

1. LOIS DISCRETES USUELLES
2. LOI DE BERNOULLI B(0,1)

Considérons une expérience dont le résultat est aléatoire et soit A un

évènement défini sur cette expérience. Tout résultat de l’expérience est une

réalisation de l’évènement A ou de son contraire Ᾱ.

Posons :

P{A} = p

P{Ᾱ} = q

Nous avons P{A} + P{Ᾱ} = 1 soit p + q = 1

Exemple :

On lance un dé. Soit A l’évènement « valeur amenée A est 6 ». Ᾱ est l’évènement «  la valeur amenée est l’un des nombres 1, 2, 3, 4, 5 » Si le dé est parfaitement équilibré, nous avons :

P{A} = p =

P{Ᾱ} = q =

Lorsqu’on s’intéresse ainsi à un évènement A ou à son contraire Ᾱ. La réalisation de l’expérience est appelée **épreuve de Bernoulli.**

On peut associer à une épreuve de Bernoulli, une variable aléatoire X prenant k valeur 1 quand l’évènement A est réalisé et la valeur 0 quant c’est l’évènement contraire Ᾱ qui est réalisé.

Loi de probabilité de X est : P{X = 1} = p

P{X = 0} = q, avec p + q = 1

Une telle variable aléatoire est appelée variable aléatoire de Bernoulli et sa loi de probabilité est appelée loi de Bernoulli.

* Espérance mathématique

E(X) = p x 1 + q x 0

E(X) = p

Le moment d’mordre 2 ou espérance mathématique de est :

E() = p x + q x

E() = p

* La variance de X est alors

= E( ) - [  
 = p -

= p(1 – p)

= pq

Exercice :

A un concours se présentent deux fois plus d’hommes que de femmes. On tire une personne au hasard et on l’appelle X la variable aléatoire « nombre de femme »

1. Quelle loi suit la variable X ?
2. Calculer E(X) et

Réponse :

1. La population étudiée est composée de 2/3 d’hommes et 1/3 de femmes.

On tire une seule personne au hasard, donc la loi de X est la loi de Bernoulli.

1. Espérance mathématique

E(X) = p P(X = 1) =

Ecart-type

= x

= 0,471

1. LOI BINOMIALE β(n,p)

Considérons une suite de n épreuves de Bernoulli identiques. A chaque

épreuve nous appelons **succès** la réalisation de l’évènement A et **échec**  la réalisation

de l’évènement contraire Ᾱ.

Soit X le nombre de réalisations de l’évènement A ou le nombre de succès au cours

de n épreuves. X est une variable aléatoire qui peut prendre les valeurs 0, 1, 2, …. n.

Sa loi de probabilité est :

P{X = k} = pour k = 0, 1, 2, 3…….n

Cette loi de probabilité est appelée loi binomiale car on reconnait en le

terme général du développement du binôme de Newton.

En effet =

Puisque p + q = 1, l’identité ci-dessus exprime simplement que la somme des

probabilités correspondant aux différentes possibles de X est égale à 1.

Notons que la loi binomiale dépend de deux paramètres n et p

Espérance mathématique et variance

E(X) = np

V(X) = npq

Remarque : Comment reconnaître une loi binomiale

On a recours à une loi binomiale si :

1. On a affaire à une expérience aléatoire comportant deux modalités complémentaires A et Ᾱ avec P{A} = p et P{Ᾱ} = q et p + q = 1
2. On réitère l’épreuve n fois .
3. Les n réalisations sont indépendantes les unes des autres
4. La valeur de p n’est pas trop voisine de 0 ni de 1

La variable aléatoire a pour valeur le nombre de succès dans la suite de n épreuves.

Exercice :

On lance un dé 5 fois consécutivement et on s’intéresse à l’arrivée d’un nombre pair.

Le dé étant parfaitement équilibré, la probabilité d’obtenir un nombre pair est, pour

épreuve, égale à =

Quelles sont les probabilités associées aux diverses valeurs de X ? Donner la représentation graphique.

Réponse :

Le nombre de fois que l’on obtient un nombre au cours de cinq épreuve est une variable aléatoire X. Sa loi de probabilité est une loi binomiale de paramètres n = 5 et p =

P{X= k} = avec k = 0, 1, ……..5

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K | 0 | 1 | 2 | 3 | 4 | 5 |
| P{X = k} |  |  |  |  |  |  |

5/32 --

1/32 --

0 1 2 3 4 5

Diagramme en bâton de la loi d probabilité β(n ;

Remarques :

1. La loi binomiale est symétrique lorsque p =  . En effet les deux termes équidistants des extrémités sont :

et

b- Lorsque p la loi binomiale devient dissymétrique. La valeur la plus probable est alors déplacée vers la gauche si p et vers la droite si p

1. LA LOI DE POISSON
2. Soit X la variable aléatoire pouvant prendre toutes les valeurs entières (variables discrètes) 0, 1, 2,…….. n avec les probabilités :

P{X = k} = avec k = 0, 1, 2,………n

où m étant un nombre positif

e la base de logarithme népérien e = 2,71828

Une telle loi de probabilité est appelée Loi de Poisson. Elle dépend de la paramètre m. On vérifie que la somme des probabilités est égale à 1.

En effet =

Alors d’après un résultat de l’analyse, la somme infinie

ou

Par conséquent

1. Les moments de la loi de Poisson P(m)

E(X) = m

V(X) = m

L’espérance mathématique et la variance de la Loi de Poisson sont égales au paramètre m de cette loi.

1. Champ d’application :

- La loi d’attente, pannes de machines, appel téléphonique dans un standard etc…

La loi de Poisson intervient dans la modélisation des phénomènes aléatoires où le futur es indépendant du passé.

- La loi de Poisson concerne les phénomènes où les conditions d’application de la loi binomiale sont réunies :

- répétition indépendante d’une même épreuve dichotomique ;

- la valeur de l’application mesurable est le nombre de succès k dans n

épreuves dont le nombre est grand.

- le produit np a une valeur finie d quelques unités…

- Théorème de convergence :

Si la variable aléatoire X suit la loi binomiale β(n,p) avec n , p

 , alors on peut remplacer la loi binomiale par la loi de Poisson de paramètre m = np

- Calcul numérique :

La loi d Poisson est un calcul relativement aisé pour les petites valeurs de k. Elle est tabulée pour certaines valeurs du paramètre m.

Exercice :

Lors d’une enquête par sondage sur un grand nombre d’individus, 2% des personnes interrogées acceptent de ne pas rester anonymes. Sachant que l’un des enquêteurs a interrogé 250 personnes, calculer la probabilité que :

1. Ces 250 personnes souhaitent rester anonymes
2. 3 personnes acceptent de ne pas rester anonymes
3. Plus de 10 personnes acceptent de ne pas rester anonymes.

Réponse :

Soit X le nombre de personnes ne souhaitant pas rester anonymes.

Après analyse, on déduit que X suit une loi binomiale β(250, 0,02)

Mais n est grand et p = 0,02, on peut approximer cette loi binomiale par loi de Poisson de paramètre np = 5

1. Par conséquent :

P{X = 0} = = 0,0067

En utilisant cette loi binomiale on aurait obtenu :

P{X = 0} =

Résultat peu différent du résultat précédent, ce qui signifie l'approximation

1. P{X = 3} =
2. P{X 10} = 1 - P{X = 1 – P{X < 11} = 1 – 0,986 = 0,140

La valeur 0,986 étant lue directement sur la table cumulée.

1. APPLROXIMATION DE LOI BINOMIALE PAR UNE LOI DE POISSON

Soit X une variable aléatoire dont la loi de probabilité est une loi de paramètres

n = 100 et p = 0,02

On a P{X = k} = pour k = 0, 1, 2,…… 100

Cette loi de probabilité est fortement dissymétrique. Les valeurs significatives sont

groupées autour de k = np = 100 x 0,02 = 2. Les probabilités correspondantes aux

valeurs élevées de k sont extrêmement faibles et pratiquement négligeables.

Comme n étant élevé et la valeur de p faible, l’expression de la loi binomiale est d’un

maniement difficile.

On montre qu’une très bonne approximation des diverses probabilités peut être

obtenue au moyen de la loi de Poisson de paramètre m = np = 2

d’où P{X = k} =

Cette approximation correspond à une propriété de la loi binomiale

On montre que :

si n → et p → 0, np = constante = m

alors

En pratique si n , p et np on peut remplacer la loi binomiale par la loi de Poisson de paramètre np.

1. LOI NORMALE OU LOI DE LAPLACE – GAUSS
2. Définition

Soit X une variable aléatoire continue pouvant prendre toutes les valeurs de à , on dira qu’elle suit un loi normale, dite loi de Gauss si sa densité de probabilité est :

f(x) =

On notera en abrégé que X suit une loi N(m,

m et sont les paramètres de cette loi et e = 2,71828

f(x)

--------------------

-------------------------

0 (m – m (m + x

1. Valeurs caractéristiques :

E(X) = m

V(X) =

Si m = 0 la loi normale est dite centrée

Si m = 0 et = 1 la loi normale est centrée réduite ou normalisée, la densité de probabilité est alors :

f(x) =

Si l’on pose T = , T est appelé variable aléatoire normale centrée réduite.

1. Importance de la distribution de Laplace-Gauss

La loi de Laplace-Gauss est importante pour plusieurs raisons :

1. Les distributions statistiques réelles s’en approchent très fréquemment.
2. Lorsque la valeur que prend un caractère a des causes très nombreuses et indépendantes les unes des autres, les valeurs de ce caractère sont normalement distribuées.

La loi de Laplace-Gauss constitue une bonne approximation des lois binomiales et de Poisson. On dit que la loi normale Laplace-Gauss est une loi limite.

La moyenne arithmétique de n variables aléatoires suivant une même loi de probabilité tend à devenir une variable de Laplace-Gauss lorsque n augmente indéfiniment.

1. Propriétés de la distribution de Laplace-Gauss.

Soit la fonction de répartition de la loi de Laplace - Gauss

Par définition

y

y =

0 t x

D’autre part : P{t1 2 } =

y

0 t1 t2

Soit Φ(t) la probabilité que x soit comprise entre 0 et t y

0 t x

La fonction de répartition est liée à par la relation :

étant la valeur de

Exercice 1

La variable aléatoire X suit la loi normale N(18,3). Calculer la probabilité de chacun des évènements suivants :

X

Solution 1

1. Si la variable aléatoire suit la loi la variable aléatoire T définie par

T = suit la loi normale centrée réduite.

Ainsi X = 3T + 18

Calculer de P{ X

D’après la table de la fonction de répartition :

P{T

P{ X

1. P{ équivaut à P{3T + 18  
   P{ = P{ T } avec

P{

D’après la table

Donc P{

1. P{15 équivaut à P{15

= P{- 1

= 1,33) – [1 –

= P{15 = 0,7495

1. P{14

= P{- 1,33

=

=

P{14

Exercice 2

La variable aléatoire X suit la loi normale de moyenne 2 et d’écart-type 0,1

Calculer le nombre h dans les cas suivants :

1. P{X
2. P{X
3. P{2 – h

Solution 2

Si X suit la loi normale N(2,0,1) alors la variable aléatoire T définie par T = suit la loi normale centrée réduite. D’où X = 0,1T + 2

1. D’où P{ X équivaut à P{ T

Cette dernière égalité montre que est négatif.

Donc

Soit la table donne

Donc

1. P{X

D’où P{ T équivaut à P

De la table de la fonction de répartition de la loi N(0 ,1) nous obtenons

P{ T d’où 0,68 = donnant 0,068 =2 – h

alors h = 1,932

1. P{2 – h

D’où {2 – h 0,1T + 2

P{ donc 2

De la table de la loi normaleN(0,1) , nous obtenons

d’où et h = 0,128

1. APPROXIMATION DE LA BINOMIALE PAR LA LOI NORMALE

La probabilité en matière de la loi binomiale est :

P{X} =

avec E(X) = np et V(X) = npq

Pour n grand, cherchons une approximation de P{X} par une loi autre que la loi de

Poisson. Faisons le changement de variable : T = et remplaçons

X par Tnp

Un raisonnement long conduit à obtenir une approximation de P{X} par la formule :

P {X} =

L’expression est dite expression de la loi normale centrée réduite.

Elle donne la densité de probabilité d’une variable aléatoire t variant de

La fonction t est tabulée.

Les conditions d’approximation de la loi binomiale β(n,p) par la loi normale N(m,

sont :

* n
* np
* npq
* p n’est ni voisine de 0 ni de 1

Exercice :

On tire un échantillon de taille n = 40 dans une population comportant une proportion p = 0,4 d’individus présentant le caractère C, par exemple posséder une voiture.

Déterminer la probabilité d’observer dans l’échantillon exactement 20 individus possédant le caractère C.

Solution :

Le nombre d’individus présentant le caractère C dans l’échantillon est une variable aléatoire binomiale

E(X) = np = 16

V(X) = npq = 9,6 et par excès

Mais comme la taille de l’échantillon est suffisamment grande n et la proportion p = 0,4 n’est voisine de 0 ni de 1, la loi binomiale peut valablement être approximée par une loi normale ayant pour paramètre m = 16 et

β(40 ; 0,4) → N(16 ; 3,1)

La variable t centrée réduite correspondant à x = 20 est :

t =

Par consultation de la table de la loi normale N(0 ; 1) et interpolation linéaire on trouve

y(t) = 0,1737 (table 5)

d’où f(x) =

La probabilité d’observer 20 individus est de 0,056

A titre de vérification, la probabilité exacte est :

P{X = 20} = =

CHAP IV – L’INTERPRETATION DES SONDAGES ALEATOIRES

INTRODUCTION :

Il s’agit de déterminer certaines caractéristiques d’une population :

- moyenne

- écart-type

- proportions d’éléments A et Ᾱ

etc….

à partir des valeurs correspondantes observées sur un échantillon extrait de cet ensemble.

On sait qu’il n’y a pas de certitude absolue quant à la précision des informations recueillies par sondage ; une part d’aléa est inhérente à cette méthode.

Le problème se pose donc à partir des observations effectuées sur l’échantillon, d’estimer avec le maximum d’efficacité la valeur des caractéristiques de la population et d’apprécier la précision de cette estimation.

A - DISTRIBUTION D’ECHANTILLONAGE DE LA MOYENNE ARTHIMETQUE

La distribution des diverses valeurs que peut prendre la moyenne

d’échantillon calculée sur tous les échantillons possibles de même taille d’une population donnée porte le nom de distribution d’échantillonnage.

D’une façon générale la distribution d’échantillonnage caractérise les fluctuations d’échantillonnage de toute statistique (moyenne, variance, proportion, etc) calculée sur tous les échantillons possibles de même taille.

La moyenne arithmétique calculée sur tous les échantillons possibles d’une population diffère d’un échantillon à un autre et va prendre diverses valeurs autour d’une valeur centrale.(moyenne de la population).

Les fluctuations de autour de cette valeur centrale seront quantifiées par une mesure de dispersion qui est l’écart-type de la moyenne arithmétique que l’on peut observer sur chaque échantillon.

Celle-ci est donc une variable aléatoire qui peut prendre diverses valeurs selon les résultats de l’échantillonnage et possédera une distribution de probabilité.

Exemple

Supposons qu’une population compte 5 éléments numérotés de 1 à 5, et le poids respectif de chacun est :

n° 1 x1 = 116 kg

n° 2 x2 = 118 kg

n° 3 x3  = 120 kg

n° 4 x4  = 122 kg

n° 5 x5  = 124 kg

Déterminez la distribution des moyennes de tous les échantillons, l’expérience d’échantillonnage étant avec remise.

La moyenne de la population est

La variance

Echantillonnage de la population avec remise

Exemple :

Supposons que nous voulons former tous les échantillons possibles de taille n = 2 de cette population en effectuant un tirage avec remise.

Dans ce cas, il y a 52  échantillons possibles. Chacun d’entre eux ayant une probabilité 1/25 d’être choisi. Les résultats possibles sont :

|  |  |  |  |
| --- | --- | --- | --- |
| Echantillon n° | Eléments | Echantillonnage | Moyenne de l’échantillon |
| 1  2  3  4  5 | 1.1  1.2  1.3  1.4  1.5 | 116 – 116  116 - 118  116 – 120  116 – 122  116 – 124 | 116  117  118  119  120 |
| 6  7  8  9  10 | 2.1  2.2  2.3  2.4  2.5 | 118 – 116  118 – 118  118 – 120  118 – 122  118 – 124 | 117  118  119  120  121 |
| 11  12  13  14  15 | 3.1  3.2  3.3  3.4  3.5 | 120 – 116  120 – 118  120 – 120  120 – 122  120 – 124 | 118  119  120  122  124 |
| 16  17  18  19  20 | 4.1  4.2  4.3  4.4  4.5 | 122 – 116  122 – 118  122 – 120  122 – 122  122 – 124 | 119  120  121  122  123 |
| 21  22  23  24  25 | 5.1  5.2  5.3  5.4  5.5 | 124 – 116  124 – 118  124 – 120  122 – 122  122 - 124 | 120  121  122  123  124 |

Les moyennes des échantillons varient entre 116 et 124 et certaines entre elles reviennent plus souvent.

La distribution de fréquences des moyennes des échantillons se présentent comme suit :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Fréquence | Fréquence relative | Ecarts  | | Carrés des écarts | Carrés des écarts x fréquence |
| 116  117  118  119  120  121  122  123  124 | 1  2  3  4  5  4  3  2  1 | 1/25  2/25  3/25  4/25  5/25  4/25  3/25  2/25  1/25 | 4  3  2  1  0  1  2  3  4 | 16  9  4  1  0  1  4  9  16 | 16/25  18/25  12/25  4/25  0  4/25  12/25  18/25  16/25 |

100 /25

Var(x) = 4

Calculons la moyenne et la variance de

E(

Par conséquent E(

Calculons maintenant l’ampleur de la dispersion des moyennes

On trouve Var(

Nous constatons que les moyennes des sont moins dispersés autour de la moyenne ( m ) de la population que les valeurs individuelles V(

Si nous continuons avec l’échantillon de taille n = 3, il y aura 125 échantillons possibles et la moyenne d’échantillonnage de restera égale 120 ; cependant la variance des moyennes de 125 échantillons va diminuer et on trouvera dans ce cas

Var( ainsi de suite

En résumé :

Si on prélève avec remise un échantillon aléatoire de taille n à partir d’une population finie dont les éléments possèdent un caractère mesurable de moyenne m et de variance alors la moyenne d’échantillon suit une loi de probabilité de moyenne E(xi) = m et de variance Var(xi) =

L’écart-type de la moyenne

Remarques

1. Lorsque l’échantillonnage est effectué sans remise, on doit apporter une correction.

Dans ce cas : Var(

Toutefois, le facteur de correction peut être ignoré si le taux de sondage est inférieur à 10%.

1. Le nombre d’échantillons possibles est
2. La moyenne d’échantillons est aussi appelée moyenne échantillon ale.

B – THEOREME CENTRAL LIMITE

Pour connaître exactement la distribution d’échantillonnage de la moyenne d’échantillons , il faut connaître la distribution de la population qui a été échantillonnée, ce qui n’est pas toujours possible ou c’est si le cas, on peut l’apparenter à une forme possible.

Cependant, il existe en statistique un théorème important qui va nous permettre de contourner cette difficulté.

1. THEOREME

Si des échantillons aléatoires de taille n sont prélevés d’une population infinie dont les éléments possèdent un caractère mesurable x de moyenne E(x) = m et de variance V(x) = , la distribution d’échantillonnage de la variable aléatoire tend alors à se rapprocher d’une loi normale de moyenne E( et de variance V( et ce, d’autant plus que la taille de l’échantillon est grande.

En pratique ce théorème s’applique dès que  , on peut déduire une valeur fiable de en calculant la variance s2  de l’échantillon tel que :

s2 = alors

1. TRANSFORMATION DE LA VARIABLE ALEATOIRE EN UNE VARIABLE CENTREE REDUITE

La variable centrée réduite s’obtient de façon suivante :

VCR =

Donc, Z =

distribuée d’après la loi normale centrée réduite lorsque est distribuée suivant la loi normale. Ainsi, la probabilité que soit comprise entre deux valeurs a et b.

Pr{

1. RESUME DES DIFFERENTES SITUATIONS CONCERNANT L’ECHANTILLONNAGE D’UNE POPULATION

Soit un échantillon de taille n d’une population dont les éléments possèdent un caractère mesurable m et

La moyenne d’échantillon est une variable aléatoire, la distribution possède les propriétés suivantes selon les caractéristiques de la population. Les cas suivants sont à distinguer :

1er cas : Population normale et variance connue

1. la distribution xi  est normale
2. la moyenne de la distribution de
3. l’écart-type de la distribution

Les fluctuations de l’écart-réduit Z =

2ème cas : La distribution de la population et la variance sont inconnues.

Dans ce cas, on utilise les résultats du théorème central limite

1. la distribution de xi est approximativement normale
2. la moyenne de la distribution de
3. l’écart-type de la distribution d’échantillonnage de est :

s2 = et qui donne une borne approximation

Les fluctuations de l’écart-réduit Z = suivent la loi normale réduite.

CHAP V LA THEORIE DE L’ESTIMATION

A partir des observations effectuées sur l’échantillon, le problème se pose d’estimer avec le maximum d’efficacité la valeur de telle ou telle caractéristique de la population :

- moyenne

- proportion

- variance etc…

Ces estimations peuvent s’exprimer soit :

- par une seule valeur, c’est l’estimation ponctuelle ;

- par intervalle, c’est l’estimation par intervalle

Ces estimations seront accompagnées d’une certaine marge d’erreur puisque l’échantillon ne donne que qu’une information partielle.

1. ESTIMATION PONCTUELLE

Estimer un paramètre par un nombre, c’est chercher une valeur la plus rapprochée

du paramètre inconnu.

Un estimateur de est une fonction qui, aux valeurs de l’échantillon, il fait correspondre

. ………..

dans lequel

Lorsqu’un paramètre inconnu d’une population statistique est estimé par un seul nombre déduit des résultats d’un échantillon, ce nombre est appelé un estimateur ponctuel du paramètre.

Remarque : l’estimateur dépend des observations d’un échantillon aléatoire. De ce fait, il est également une variable aléatoire.

Propriétés d’un estimateur

1. Estimateur sans biais ou sans dispersion

étant un paramètre de valeur inconnue et  l’estimateur de

Un estimateur est sans biais si la moyenne de sa distribution d’échantillonnage est

égale à la valeur du paramètre de la population à estimer, c'est-à-dire si :

E( ) =

Si l’estimateur est biaisé, son biais est mesuré par :

Biais = E() -

1. Estimateur convergent

Un estimateur  est convergent si la distribution tend à se concentrer autour de la

valeur inconnue à estimer à mesure que la taille de l’échantillon augmente, c'est-à-dire si

V( à mesure que n tend vers l’infini. est un estimateur convergent de m puisque

lim V(Remarques :

* Un estimateur sans biais et convergent est absolument correct
* Un estimateur qui est asymptotiquement sans biais et convergent est dit estimateur correct.

1. Estimateur efficace

Un estimateur sans biais est plus efficace (ou simplement efficace) si sa variance est l

la plus faible parmi les autres variances des autres estimateurs sans bais.

Ainsi, sont deux estimateurs sans biais du paramètre , l’estimateur est plus efficace si :

1. ESTIMATION PAR INTERVALLE

Estimer par intervalle un paramètre consiste à trouver deux nombres a et b telle

la probabilité de à estimer soit :

Pr{

( est le seuil de confiance ou d’acceptation (choisi à priori). L’intervalle (a , b) est l’intervalle de confiance.

étant le coefficient de risque.

Cet intervalle de confiance signifie que, si nous répétons un grand nombre de fois un échantillon de taille n de la même population, par exemple dans 100 (1 – cas sur 100, l’intervalle recouvre la valeur du paramètre.

Pour déterminer cet intervalle, l’on doit connaître la distribution d’échantillonnage (distribution de probabilité) de l’estimateur correspondant, c'est-à-dire connaître la façon dont sont distribuées toutes les valeurs possibles de l’estimateur obtenues à partir de tous les échantillons possibles de taille prélevés dans l population.

2-1 Estimateur d’une moyenne par intervalle de confiance

On propose d’estimer, par intervalle de confiance, la moyenne m d’une population l’écart-type est

On tire au hasard dans la population un échantillon de n unités et soit la moyenne correspondante.

On répète un grand nombre de fois le tirage de n unités

On montre que la moyenne de l’échantillon est une variable aléatoire qui suit une loi de Laplace GAUSS de moyenne arithmétique m et d’écart-type dès que l’effectif de l’échantillon dépasse une trentaine d’unités.

Si la distribution du caractère mesurable est inconnue ou si la variance de la population est inconnue, un échantillon de taille n nous permet, d’après le théorème central limite, de considérer que suit approximativement un loi normale.

La probabilité que soit compris dans l’intervalle [m - t

Nous pouvons écrire x -

* Dans le cas d’un échantillon tiré avec remise
* Dans le cas d’un échantillon exhaustif

On a donc la possibilité, avec une probabilité fixée et à partir d’un échantillon, de déterminer un intervalle dans lequel la moyenne doit se trouver, la possibilité d’estimer la moyenne arithmétique.

Quand l’écart-type de la population est connu, le calcul ne présente pas de difficulté.

Mais, quand l’écart-type n’est pas connu, on prendra pour valeur approchée l’écart-type calculé sur l’échantillon.

Exemple :

Sur un échantillon de 100 pièces industrielles, on a relevé une côte déterminée x. On a trouvé une moyenne avec un écart-type

1. Estimer la moyenne vraie m par un intervalle de confiance à 95% , à 98%, 99,7%
2. Quel devrait être l’effectif de l’échantillon pour situer la moyenne m dans un intervalle de 0,6 mm avec un sécurité de 98% ?

Solution :

1. est estimé par

- Les limites de l’intervalle de confiance à 95% sont données par :

On a donc 49,69mm

- Les limites de l’intervalle de confiance à 98% sont :

On a ainsi 49,63mm

- Les limites de l’intervalle de confiance à 99,7% sont :

Soit 49,51mm

1. Pour situer la moyenne m dans un intervalle de 0,6mm avec une sécurité de 98%,

l’effectif de l’échantillon doit vérifier l’égalité suivante :

Soit 2,33 x

d’où n = 155

2-2 Estimation d’une proportion par intervalle de confiance

On se propose d’estimer la proportion des individus possédant le caractère C, une proportion qui est inconnue. Prélevons un échantillon de n individus dans lequel on détermine la proportion des individus possédant le caractère C, proportion pe  = . Le pourcentage pe  = est donc une variable aléatoire, il varie suivant l’échantillon prélevé de façon aléatoire. Les pourcentages p1, p2,…………,pn  constituent une série statistique appelée distribution d’échantillonnage des pourcentages.

Puisque la seule information dont nous disposons résulte de l’observation de l’échantillon, la meilleure estimation de p est p = pe.

Nous savons également que la distribution d’échantillonnage du pourcentage

p = est une loi binomiale de moyenne p et d’écart-type

Pour n assez grand, n et p ni trop voisin ni de 0 ni de 1, la loi binomiale peut valablement être approximée par la loi normale de même moyenne et de même écart-type.

Si l’on fixe un seuil de confiance 1 – , il existe un réel unique

tel que Pr{

Comme p est inconnu et intervient au niveau des bornes de son propre encadrement :

1. La première méthode consiste à remplacer p par pe. Mais comme est un écart-type, il faut le multiplier par et nous obtenons :

L’intervalle est l’intervalle de confiance de p au seuil de confiance ou au coefficient de risque

1. Une autre méthode consiste à utiliser le résultat suivant : sur l’intervalle de probabilité [0 ; 1] le produit p(p – 1) est maximal si

En majorant on obtient comme intervalle

L’élargissement de cet intervalle n’est considéré comme acceptable que si pe est comprise entre 0,3 et 0,7. Dans le cas particulier où que l’on peut majorer par 2.

Si pe est comprise entre 0,3 et 0,7 un intervalle de confiance p au risque

Exemple :

On considère une urne contenant des boules identiques mais de couleurs différentes.

On tire au hasard une boule de l’urne, on note sa couleur et on la remet dans l’urne. Sachant que une suite de 400 tirages, on a obtenu 88 fois une boule blanche, donnez une estimation de la proportion des boules blanches dans l’urne et l’intervalle de confiance correspondant aux coefficients de risque 2% et de 5%

Solution

Le pourcentage de boules blanches dans l’échantillon étudié est :

p =

Puisque la seule information disponible résulte de l’observation, la meilleure estimation de p0  est p.

p0  = p = 0,22

Comme n et que p0  n’est pas trop voisin ni de 0 ni de 1, la répartition d’échantillonnage du pourcentage est très proche d’une loi normale de moyenne p0 et de l’écart- type

Nous avons ainsi (1 – ) % chances de ne pas nous tromper en affirmant que la valeur exacte de p0 est comprise dans l’intervalle :

Or

L’intervalle de confiance est donc :

1. pour

Soit 0,22 –  ; 0,22 + (2,326 x 0,0207)

L’intervalle est donc (0,1718 ; 0,268)

1. pour

Soit 0,22 – ; 0,22 + (1,96 x 0,0207)

L’intervalle est donc (0,1794 ; 0,26)

CHAP VI LA THEORIE DES TESTS D’HYPOTHESE

INTRODUCTION :

L’information du statisticien, ne porte que sur un nombre limité de valeurs qui composent l’échantillon. C’est la méthode de l’estimation qui consiste à fournir des réponses générales à partir d’informations partielles.

Bien des problèmes dans la pratique ne se présentent pas toujours en termes d’estimation mais en termes de comparaison, comparaison d’une côte obtenue à partir d’un sondage aléatoire à une norme fixée à priori ou encore à comparer entre eux les résultats de deux échantillons différents.

La résolution de ces problèmes de comparaison à partir d’échantillons aléatoires repose sur un mode de raisonnement statistique désigné sous le nom de TEST D’HYPOTHESE

1. Concepts importants dans l’élaboration d’un test d’hypothèse

Définitions

1. Hypothèse statistique : Une hypothèse statistique est un énoncé, une affirmation concernant les paramètres d’une population.
2. Test d’hypothèse : Un test d’hypothèse ou test statistique est une démarche qui a pour but de fournir une règle de décision permettant sur la base de résultats d’échantillon, de faire un choix entre deux hypothèses statistiques.

Les hypothèses statistiques qui sont envisagés à priori s’appellent :

- l’hypothèse nulle notée H0

- l’hypothèse alternative dite H1

Un des aspects importants d’un test d’hypothèse est de convenir d’avance à quelle condition l’une ou l’autre des hypothèses sera considéré comme vraisemblable.

Hypothèse nulle H0, l’hypothèse selon laquelle on fixe à priori un paramètre de la population à une valeur particulière s’appelle l’hypothèse nulle H0.

N’importe quelle autre hypothèse qui diffère de l’hypothèse H0 s’appelle l’hypothèse alternative ou contre-hypothèse notée H1.

C’est l’hypothèse nulle qui est soumise au test et toute la démarche du test s’effectue en considérant comme vraie.

Si le test conduit, d’après le résultat de l’échantillon, au rejet de l’hypothèse nulle, nous considérons alors l’hypothèse alternative H1 comme vraisemblable plutôt que H0.

Notons que, quelle que soit notre décision, elle peut être erronée :

* soit en refusant (H0) alors qu’elle est vraie. Ce risque d’erreur s’appelle risque de première espèce et est généralement noté
* soit en acceptant (H0) alors qu’elle est fausse. Ce risque d’erreur est appelé risque de seconde espèce et est généralement noté

1. Seuil de signification d’un test d’hypothèse

Le risque consenti à l’avance et qui est noté , de rejeter à tort l’hypothèse nulle

H0 alors qu’elle est vraie( et de favoriser alors l’hypothèse alternative H1 ) seuil de signification du test et s’énonce en probabilité comme suit :

Le tableau suivant résume la situation

|  |  |  |
| --- | --- | --- |
| **H0**  **Décision** | **Vraie** | **Fausse** |
| Accepter | Décision correcte(risque 1 –) | Erreur de seconde espèce (risque ) |
| Refuser | Erreur de première espèce( risque ) | Décision correcte risque(1 – ) |

Exemple :

Un marchand livre à un client des lots d’objets. Un contrat fixe la proportion maximale d’objets défectueux que comporter chaque lot :

Par rapport à ce contrat, le service de réception du client peut être amené à commettre deux types d’erreurs.

|  |  |  |
| --- | --- | --- |
| Lots  Décision | Corrects | Défectueux |
| Accepter | Décision correcte | Erreur de seconde espèce :  Risque de l’acheteur |
| Refuser | Erreur de première espèce :  Risque du commerçant | Décision correcte |

Remarque :

La conclusion déduite des résultats de l’échantillon suivant la règle de décision qu’on aura adoptée aura un caractère probabiliste. La prise de décision entraîne un certain risque qu’elle soit erronée. Ce risque est donné par le seuil de signification du test.

A ce seuil de signification, on fait correspondre sur la distribution d’échantillonnage de la statistique ou sur celle de l’écart-réduit, une région de rejet de l’hypothèse nulle appelée également région critique. L’aire de cette région correspond à la probabilité . Cette région de rejet H0 est constituée d’un ensemble des valeurs de la statistique qui conduiront au rejet H0.

Sur la distribution d’échantillonnage correspondra aussi une région complémentaire dite région de non rejet de H0 appelée également région d’acceptation 1 -

La valeur observée de la statistique déduite des résultats de l’échantillon appartient soit à la région de rejet de H0, soit à la région de non rejet de H0 (on favorise alors l’hypothèse H1)

COMPARAISON A UN STANDARD

Le problème de comparaison d’une grandeur estimée à partir d’un échantillon à une valeur fixée à priori (standard, norme, etc…) est très fréquent. Exemples : pourcentage d’erreurs ou déchets, valeur moyenne d’un composant électronique etc… sont-ils égaux aux valeurs spécifiées ?

Ce problème revient au test de deux hypothèses alternatives H0  et H1

L’hypothèse H1 peut prendre des formes différentes suivant la nature de la question posée, étant comparée à un standard

Chacun de ces 3 cas à des règles de tests différents : dans le premier la région critique est tout entière à droite de l’intervalle de variation de  ; dans le second entièrement à gauche, dans le troisième systématiquement à droite et à gauche de l’intervalle de variation.

1. Test relatif à une moyenne

Prélevons un échantillon de taille n de la variable X. Sa moyenne échantillonnale

étant . Considérons la moyenne de X pour toute la population soit m. En raison des fluctuations d’échantillonnage peut différer de m. Proposons-nous de tester, sur la base observée, si la moyenne m peut être considérée comme égale à une valeur m0 fixée à priori.

En fonction du problème posé, on définit les deux hypothèses alternatives H0  et H1, qui selon les cas pourront être :

Si la population est distribuée normalement de variance connue ou si l’effectif de l’échantillon est suffisamment grand, la moyenne suit une loi normale N(m ;.

Supposons que l’hypothèse H0(m = m0) soit exacte, la variable :

suit une loi N(0,1)

Etant donné le seuil de signification , on peut déterminer la région correspondant à chacun des 3cas précédents.

Ainsi, dans le 3ème test d’hypothèse la région d’acceptation est de la forme

Les valeurs de étant déterminées de façon à ce que la probabilité

Pr{choisir H0/H0 vraie} = Pr{

La région d’acceptation est donc définie par :

où est la valeur de la variable N(0,1)

tel que :

l1  m0 l2

Cela signifie que, si l’hypothèse H0 est vraie, le risque de prélever un échantillon aléatoire de taille n tel que est au plus égal à (seuil de risque fixé).

La règle de décision du test en résulte :

- on rejette H0 avec le risque de se tromper

- on n’a aucune raison de rejeter H0 donc on l’accepte mais on a encore un risque de seconde espèce non quantifiée de se tromper.

Remarque

Si l’écart-type est inconnu, on remplace

puis on calcule et on applique la règle de décision précédente.

Exercice 1

Une machine produit des pièces dont la moyenne des diamètres est de m0 = 50 et de . Seuil de risque retenu 5%. Sur un échantillon de 100 pièces la moyenne est de 50,1.

Alors

On a donc

En conclusion, la machine est déréglée avec le risque de se tromper.

Exercice 2

Une machine fabrique des pièces métalliques en série. Elle a été réglée pour que le diamètre de celle-ci soit égal à 12,60mm, mais une certaine variabilité est inévitable. Sur un échantillon de 100 pièces, on a observé une moyenne .

Le réglage de la machine peut-il encore être considéré comme correct ?

Seuil de signification .

Solution :

Dans cet exemple, on se propose de tester l’hypothèse :

La taille de l’échantillon étant suffisamment grande pour que la moyenne d’échantillon suive une loi normale

La vraie valeur de est inconnue, mais on l’estimer valablement par s2  tel que :

s = 0,40

En supposant l’hypothèse H0 exacte, la variable T :

est distribuée normalement.

En raison des hypothèses alternatives retenues, la région d’acceptation sera de la forme :

Pour le seuil de signification , la valeur de la variable normée est telle que :

La région d’acceptation sera donc :

La valeur observée se trouvant à l’intérieur de cette région d’acceptation n’est pas contradiction avec l’hypothèse H0.

Les mesures effectuées sur l’échantillon ne mettent pas en doute le réglage de la machine.

1. Test sur une proportion

Dans cette section, nous nous proposons de tester si la proportion p d’éléments dans la population qui présentent un certain caractère qualitatif peut être considérée ou non comme égale à une valeur hypothétique p0.

La proportion p d’individus dans la population est inconnue. La proportion pe observée dans l’échantillon aléatoire peut être différer en raison de fluctuations d’échantillonnage.

Le tableau suivant précise les conditions du test

|  |  |
| --- | --- |
| TEST RELATIF A UNE PROPORTION  Conditions d’application : Echantillon de grande taille prélevé au hasard d’une population binomiale de sorte que .  Taille des cas n est suffisante  Hypothèse nulle H0: p = pe  Seuil de signification :  Ecart-réduit et sa distribution :  En supposant H0 vraie et selon les conditions d’application, l’écart-réduit  est distribuée suivant N(0,1) | |
| **Hypothèses alternatives** | **Règle de décision** |
|  |  |
|  |  |
|  |  |

On pourrait également préciser les règles de décision en fonction des valeurs critiques de p soient :

|  |  |
| --- | --- |
| **Hypothèses alternatives** | **Règle de décision** |
|  |  |
|  |  |
|  |  |

Exercice d’application :

Aux dernières élections un parti politique a obtenu 42% des suffrages. Au dernier sondage effectué par un cabinet d’études, on a obtenu 458 voix en faveur de ce parti sur 1041 électeurs.

Le Chef de parti déclara que la popularité de son parti était à la hausse. Au seuil de signification que penser de cette affirmation ?

Réponse

1. Hypothèses statistiques :

H0 : p = 0,42

H1 : p

1. Seuil de signification
2. Conditions d’application : il faut que

On a 1041 x 0,42 = 437,22

1041 x 0,58 = 603 78

Ce qui est suffisant

1. La statistique qui convient pour le test est pe estimateur de p :

L’écart-réduit est, selon les conditions d’application et en supposant H0  vraie.

p0 = 0,42 variable aléatoire distribuée suivant une loi N(0,1)

1. Règle de décision :

De façon évidente, le risque n’est pas réparti de façon symétrique

La zone d’acceptation est à gauche.