STATISTIQUES INFERENTIELLES

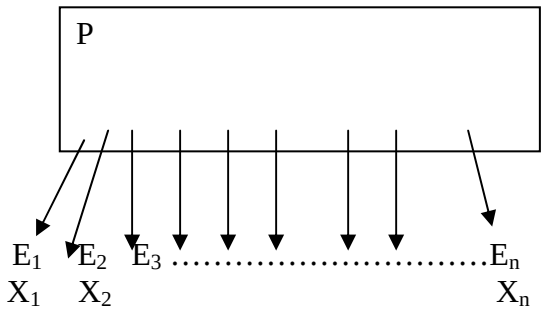
I Introduction. Principe de la théorie.

Pour étudier une population statistique il y a deux méthodes :

- La méthode exhaustive ou recensement : on examine chacun des éléments de la population (méthode trop longue)
- La méthode des sondages : on examine une partie de la population ou encore un échantillon pour essayer d'en déduire des informations sur la totalité de la population. Cette méthode comporte deux étapes :
    - L'échantillonnage qui permet de sélectionner un échantillon de la population.
    - L'estimation qui grâce aux résultats obtenus sur l'échantillon permet d'induire des Informations sur la population totale.

Principe de la théorie

On considère une population P d'effectif N dont on étudie un caractère.
Pour ce caractère on suppose que P a une moyenne m et un écart – type σ.
De cette population on extrait n échantillons à partir desquels on définit n variables aléatoires indépendantes $X_1, X_2 , \dots , X_n$ qui ont une toutes la même moyenne m et le même écart – type σ



On définit alors la variable aléatoire « moyenne » définie par $\bar{X} = (X_1 + X_2 + \dots + X_n) /n$

Théorème de la limite centrée

Si $n > 30$ la variable aléatoire $\bar{X}$ suit approximativement la loi normale $N (m , \frac{\sigma}{n})$

Lorsqu'on utilise ce théorème on dit qu'on se situe dans le cadre de la théorie des grands échantillons.

Ce qui veut dire que la variable aléatoire $\frac{\bar{X} - m}{\sigma/\sqrt{n}}$ suit approximativement la loi normale N(0,1)

II ESTIMATIONS

1°) Estimations ponctuelles

a) D'une fréquence

On a une population dont on veut étudier un caractère. (Par exemple savoir dans une usine qui fabrique une pièce pour un jouet la fréquence des pièces non-conformes)

Lorsque l'on ne connaît pas la fréquence p de ce caractère pour la population toute entière, on prélève au hasard un échantillon de cette population, on détermine la fréquence f du caractère pour cet échantillon et on considère que la fréquence trouvée peut – être affectée à la population toute entière .
(Par exemple si sur 100 pièces prélevées au hasard il y en a 4, f = 0,04)

Cette fréquence f constitue une estimation ponctuelle p de la fréquence du caractère et on a : p = f

Exercice extrait d'un sujet de BTS

On considère un échantillon de 100 chaudières prélevées au hasard dans un stock. Ce stock est assez important pour qu'on puisse assimiler ce tirage à un tirage avec remise.
On constate que 94 chaudières sont sans aucun défaut.

1. **Donner une estimation ponctuelle de la fréquence inconnue p des chaudières de ce stock qui sont sans défaut.**

b) D'une moyenne et d'un écart-type.

Si l'échantillon choisi a une moyenne $\bar{x}$ et un écart – type $\bar{\sigma}$

on considère que la moyenne m de la population est $m = \bar{x}$

et que l'écart – type σ est donnée par la formule $\sigma = \sqrt{(n/n-1)} \times \bar{\sigma}$.

2°) Estimations par intervalle de confiance

a) D'une moyenne lorsque l'on connaît l'écart- type

Soit un réel a >0.
On considère une population dont le caractère étudié est
UNE VARIABLE ALEATOIRE X QUI SUIT LA LOI NORMALE N(m, σ)
avec m INCONNUE ET σ CONNUE.

Soit $\bar{x}$ la moyenne d'un échantillon de taille n extrait de cette population alors

l'intervalle $[\bar{x} - a\,\sigma/\sqrt{n} ; \bar{x} + a\,\sigma/\sqrt{n}]$ est appelé intervalle de confiance de m au niveau ou au seuil de confiance de $2\,\Pi(a) - 1$ ce qui veut dire que il y a $2\,\Pi(a) - 1$ de « certitude » que la moyenne se situe dans cet intervalle.

Généralement on décide d'un seuil de confiance à l'avance, le plus souvent 0,95 ou 0,99 et on cherche le réel a positif tel que $2\,\Pi(a) - 1$ soit égal à ce coefficient de confiance.

Remarques : les valeurs usuelles du niveau de confiance sont 0,95 ou 0,99.

On dit aussi au risque de 5% ou de 1%.

Si c'est 0,95 alors la table nous donne a ≈ 1,96.

Si c'est 0,99 alors la table nous donne a ≈ 2,575.

b) **D'une fréquence**

Soit une population dont le caractère étudié a une fréquence inconnue p.
Soit un réel a>0. On prélève un échantillon de taille n.
On se place dans le cas où l'on peut considérer que la variable aléatoire suit la loi normale

$N(p, \sqrt{p(1 – p) /n})$.

Si f est la fréquence d'un échantillon de taille n extrait de la population alors

L'intervalle $[f – a\sqrt{f(1 – f) / n-1} ; f + a\sqrt{f(1 – f) / n-1}]$ est appelé intervalle de confiance

de p au niveau de confiance de $2 \Pi(a) – 1$.

Mêmes remarques que pour la moyenne.

Exercice extrait d'un sujet de BTS

On considère un échantillon de 100 chaudières prélevées au hasard dans un stock. Ce stock est assez important pour qu'on puisse assimiler ce tirage à un tirage avec remise.
On constate que 94 chaudières sont sans aucun défaut.

1.Donner une estimation ponctuelle de la fréquence inconnue p des chaudières de ce stock qui sont sans défaut.

2. Soit F la variable aléatoire qui à tout échantillon de 100 chaudières prélevées au hasard et avec remise dans ce stock , associe la fréquence des chaudières de cet échantillon qui sont sans aucun défaut.

On suppose que F suit la loi normale de moyenne p et d'écart – type $\sqrt{ p(1 – p) /100}$

Où p est la fréquence inconnue des chaudières du stock qui sont sans aucun défaut.

Déterminer un intervalle de confiance de la fréquence p avec le coefficient de confiance de 95%. Arrondir les bornes à 10^{-2}.

f = 0,94, n = 100 et $2 \Pi(a) – 1 = 0,95$ donc a = 1, 96. Alors I = [0,89 ; 0,99].

III) TEST DE VALIDITE D'HYPOTHESE

A notre époque on assiste à l'influence de plus en plus grande des statistiques sur les prises de décisions(on a ou on n'a pas le droit de…) en particulier dans le domaine économique.

Afin de prendre des décisions concernant toute une population on étudie un ou plusieurs échantillons de la population en émettant des hypothèses qui peuvent confirmées ou rejetées grâce à un test et qui induiront les mesures à prendre pour l'ensemble de la population.

Il s'agit d'un test de validité d'hypothèse.

1)Test relatif à une moyenne. Etude d'un exemple.

Une fois déterminée la variable aléatoire X de décision, dont la moyenne est m et qui suit la loi N(m , σ),
on considère la variable aléatoire moyenne $\overline{X}$ associée à des échantillons de taille n qui suit alors approximativement la loi N(m, σ /√n).

Exemple : Une machine produit des rondelles dont l'épaisseur est une variable aléatoire X D'écart – type 0,3 mm. La machine a été réglée pour obtenir des épaisseurs de 5 mm.
Un contrôle portant sur un échantillon de 100 rondelles a donné 5, 07 mm comme moyenne des épaisseurs de ces 100 rondelles.
Peut – on affirmer que la machine est bien réglée ?

Soit m la moyenne des épaisseurs de toutes les rondelles produites par la machine ainsi réglée c'est – à – dire la moyenne de X.

$\overline{X}$ suit donc la loi N(m , 0,3/√100) càd N(m, 0,03) (n = 100 ici).

On en déduit **la construction** d'un test de validité d'hypothèse qui nécessite les 3 étapes suivantes :

Etape 1 : On choisit l'hypothèse nulle H_0 , généralement on décide que $H_0 : m = m_0$
Où m_0 est la moyenne que l'on devrait obtenir.
C'est donc l'hypothèse que l'on va tester.

$\overline{X}$ suit alors la loi N(m_0 , σ /√n)

Exemple : dans notre exemple $H_0 : m = 5$. Donc si H_0 est vraie $\overline{X}$ suit la loi N(5 ; 0,03).

l'hypothèse H_1 est l'hypothèse alternative celle que l'on acceptera dans le cas où l'hypothèse nulle serait rejetée.

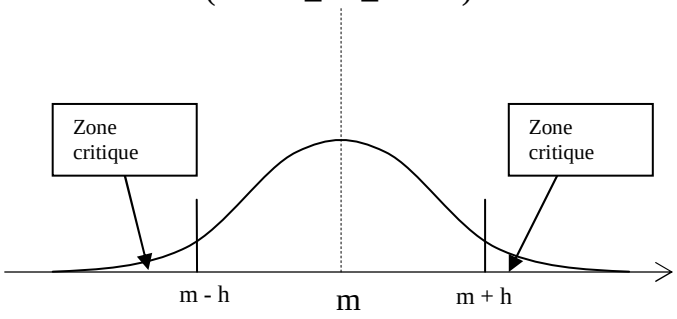
Exemple : Dans notre exemple $H_1 : m \neq 5$.

Etape 2 : L'hypothèse nulle étant considérée comme vraie on détermine la région critique à un

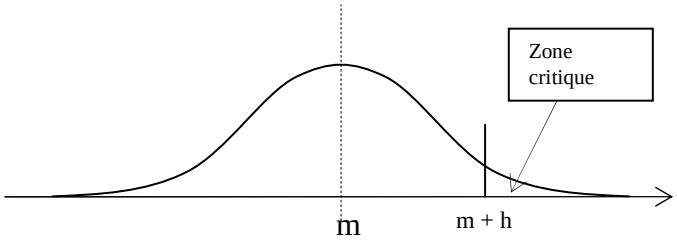
seuil α donné en admettant que la variable aléatoire $\overline{X}$ suit la loi normale N (m , σ).
Attention si le seuil de risque est de α, 1 – α est le niveau de confiance.

Trois possibilités se présentent : on cherche h tel que

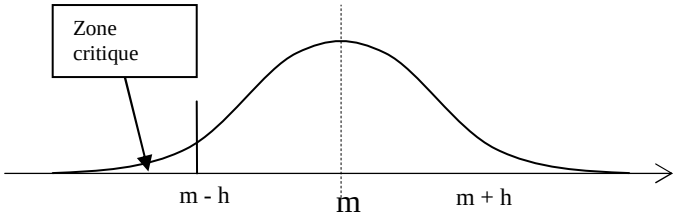
- Bilatéral : $P(m – h \leq \bar{X} \leq m + h) = 1 - \alpha$ d'où la zone critique qui est située à l'extérieur de cet intervalle.



Unilatéral : $P(\bar{X} \leq m + h) = 1 - \alpha$ d'où zone critique



ou $P(\bar{X} \geq m - h) = 1 - \alpha$ d'où zone critique



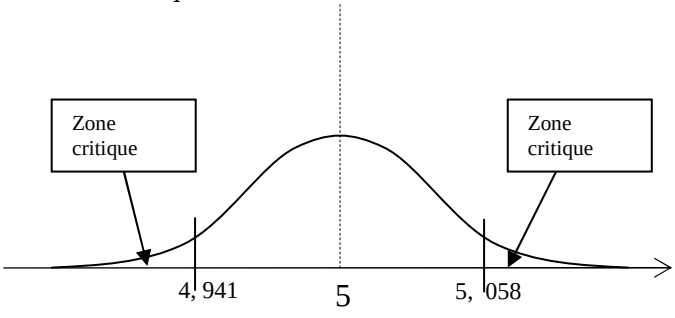
On cherche donc h tel que $P(5 – h \leq \bar{X} \leq 5 + h) = 0,95$.

Ce qui donne $P\left(\frac{-h}{0,03} \leq \frac{\bar{X} - 5}{0\ ,03} \leq \frac{h}{0,03} \right) = 0,95$ soit $2\ \Pi\ (h/0,03) - 1 = 0,95$

d'où $h/0\ ,03 = 1,96$ et $h = 0,0588$.

L'intervalle de confiance est [4,941 ; 5,058], il y a 5% de risque de ne pas avoir m dans cet intervalle.

La zone critique est la zone située à l'extérieur de cet intervalle.



Etape 3 : On énonce la règle de décision à savoir que si un ou des paramètres sont dans la zone critique on rejette H_0.

Exemple : On décide que si m dans [4,941 ; 5,058] on accepte H_0 sinon on rejette H_0.Ensuite on peut passer à la phase **d'utilisation du test**.

phase 1 : On calcule le paramètre du (ou des échantillons).

Exemple : Dans notre exemple il est donné et vaut 5,07 mm.

phase 2 : On applique la règle de décision.

Exemple : On décide donc que la machine n'est pas bien réglée et de rejeter l'hypothèse nulle.
En effet ici la moyenne de l'échantillon est dans la zone critique puisque elle est de 5,07 mm.
Et on a un risque de 5% de se tromper, en effet il se peut que l'échantillon choisi ne soit
Pas « un bon » échantillon càd qui n'est pas représentatif de la réalité.

2)Test relatif à une fréquence. Etude d'un exemple.

Exemple :
Dans un jeu de hasard qui consiste à choisir des cartes parmi 32, le tirage de certaines cartes permet un gain plus ou moins important. On veut vérifier si un joueur triche ou pas car sur 800 essais il a tiré 134 fois un as qui est la carte qui donne le plus d'avantages et ceci au seuil de risque de 1% (donc au niveau de confiance de 99%).

Soit Z la variable aléatoire qui à chaque échantillon de 800 essais, associe la fréquence d'apparition d'un as. On considère que la taille des échantillons est suffisamment grande et que

F suit la loi N(p, $\sqrt{(p(1-p)/800)}$. F est la variable aléatoire de décision.

Etape 1 : choix des hypothèses
La probabilité d'apparition d'un as est 4/32, donc si le joueur n'est pas un tricheur p= 0,125.
H_0 : p = 0,125 (le joueur n'est pas un tricheur)
Si p < 0, 125 le joueur n'est pas un tricheur non plus donc l'hypothèse alternative H_1 est
p> 0,125. (le joueur est un tricheur).

Etape 2 : Détermination de la zone critique

Dans le cas où H_0 est vraie F suit la loi N(0,125 ; $\sqrt{(0,125\times0,875/800)}$) soit N(0,125 ; 0,0117).
On cherche h tel que P(F $\geq$ 0,125 + h) = 0,01 donne h $\approx$ 0,027.

Donc la zone critique est] 0,152 ; + ∞ [.

Etape 3 : Règle de décision

Si p < 0,152 on accepte H_0 sinon on rejette H_0 .

UTILISATION DU TEST : L'échantillon observé a une fréquence égale à 134/800 soit 0,1675.
Or 0,1675> 0,152 donc H_0 n'est pas validée et le joueur est un tricheur.

3) Test de comparaison de deux moyennes

Il s'agit de déterminer s'il y a uns différence significative entre les caractéristiques de deux populations.

Si on a deux variables aléatoires indépendantes $\overline{X}_1$ et $\overline{X}_2$ d'échantillons de deux populations qui suivent respectivement les lois normales $N(m_1 ; \sigma_1)$ et $N(m_2 ; \sigma_2)$.
Alors la variable aléatoire

$D = \overline{X}_1 - \overline{X}_2$ suit la loi normale $N(m_1 - m_2 ; \sqrt{(\sigma_1^2 + \sigma_2^2)})$