NATURE DU PROBLEME.

D'une manière générale, il s'agit, à partir de l'étude d'un ou de plusieurs échantillons, de prendre des décisions concernant l'ensemble de la population.
La procédure qui consiste à préciser:

- comment un ou plusieurs échantillons doivent être prélevés dans la population étudiée
- quelles mesures doivent être effectuées sur ce ou ces échantillons
- quelle décision doit être prise à propos de l'ensemble de la population étudiée, suivant les résultats obtenus sur le ou les échantillons

s'appelle en statistique un test de validité d'hypothèse.

LES HYPOTHESES ET LES ERREURS.

Considérons une population donnée. On désire que la moyenne de cette population, lorsqu'on l'observe suivant un certain caractère, ait une valeur m_0 donnée.
On suppose que la différence entre la moyenne souhaitée (m_0) et la moyenne effective de la population (m) est nulle: c'est l'hypothèse nulle, notée $H_0 : m = m_0$.

On sait que la distribution des moyennes suit la loi normale N (m , $\frac{\sigma}{\sqrt{n}}$), c'est à dire que la variable aléatoire $\overline{X}$, qui, à tout échantillon aléatoire non exhaustif de taille n, associe la moyenne de cet échantillon, suit approximativement la loi normale N (m , $\frac{\sigma}{\sqrt{n}}$), où m est la moyenne, n l'effectif de chaque échantillon et σ l'écart-type, donc $\frac{\sqrt{n}}{\sigma}$ ($\overline{X}$ - m) suit la loi normale centrée réduite N (0 , 1).

En utilisant la loi normale centrée réduite, on peut définir une valeur de t qui permette d'obtenir $p(-t \le T \le t) = 1-\alpha$ et donc un intervalle de confiance avec une probabilité donnée.
On prélève alors un échantillon de taille n et on calcule sa moyenne $\bar{x}$
Si $\bar{x}$ appartient à l'intervalle défini on accepte H_0, sinon on rejette H_0.

Exemple:

Une entreprise reçoit un lot de 500 pièces. Le fournisseur indique que ces 500 pièces ont une masse moyenne de 780 g, l'écart-type étant de 12,5 g.
L'entreprise décide de faire un test, de garder le lot si le test est favorable, de le renvoyer au fournisseur dans le cas contraire, avec un risque de 5 %.
L'hypothèse nulle, notée H_0 est m = 780. On prélève un échantillon de taille n = 36.

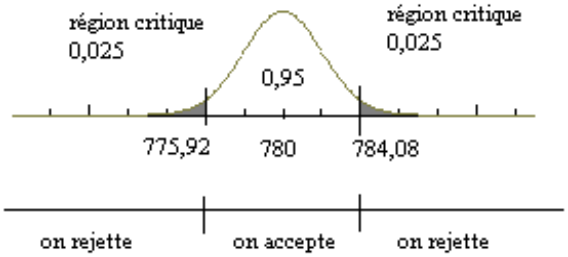
D'après la relation p ($m - t \frac{\sigma}{\sqrt{n}} \le \overline{X} \le m + t \frac{\sigma}{\sqrt{n}}$) = $1-\alpha$ on obtient p ($775{,}92 \le \overline{X} \le 784{,}08$) = 0,95

On sait donc, avant de prélever un échantillon de taille n = 36, que sa moyenne appartient à l'intervalle [775,92 ; 784,08] avec la probabilité 0,95.
Autrement dit, si l'hypothèse H_0 est vraie, il n'y a que 5% de chances de prélever un échantillon aléatoire de taille n = 36 dont la moyenne soit inférieure à 775,92 ou supérieure à 784,08.

Soit $\bar{x}_e$ la moyenne de l'échantillon prélevé:

- Si $\bar{x}_e \in$ [775,92 ; 784,08], on accepte H_0 donc on considère que la moyenne de la population est 780, le lot est gardé.
- Si $\bar{x}_e \notin$ [775,92 ; 784,08], on rejette H_0 donc on considère que la moyenne de la population n'est pas 780, le lot est renvoyé.

Dans cet exemple, si H_0 est vraie, on prend le risque de se tromper dans 5 % des cas, en rejetant à tort H_0 .
On définit ainsi une région critique au seuil α = 5 %. Le seuil α est la probabilité de rejeter H_0 alors que H_0 est vraie. Il correspond à l'erreur de première espèce. En général on fixe à priori la valeur de α .



Une autre erreur consiste à accepter H_0 alors que H_0 est fausse: c'est l'erreur de seconde espèce, dont la probabilité est notée β.
L'idéal serait de rendre les nombres α et β les plus petits possibles mais, *n* étant fixé, diminuer l'un c'est augmenter l'autre. La seule façon de diminuer les deux c'est d'augmenter *n*, ce qui n'est pas toujours possible.
En fait, la plupart du temps, les erreurs des deux types n'ont pas la même importance et on essaie de limiter la plus grave.

$1-\beta$ est la probabilité de rejeter H_0 alors que H_0 est fausse: c'est, par définition, la puissance du test.

TEST BILATÉRAL, TEST UNILATÉRAL.

Un test est bilatéral quand la région critique est située des deux côtés de la région où l'on accepte H_0.
L'hypothèse alternative notée H_1 se traduit alors par $m \neq m_0$.
On peut aussi rencontrer des situations où l'on choisit par exemple $m < m_0$ comme hypothèse alternative H_1. Le test est alors unilatéral et la région critique est alors entièrement située d'un côté de la région où l'on accepte H_0.

DÉMARCHE.

En général, les questions faisant intervenir un test de validité d'hypothèse peuvent être résolues en adoptant le plan suivant:

A . Construction du test.
1. Choix de l'hypothèse nulle H_0 et de l' hypothèse alternative H_1
2. Détermination de la région critique à un seuil α donné
3. Enoncé de la règle de décision: Si un paramètre de l'échantillon est dans la région critique, on rejette H_0, sinon on accepte H_0.

B . Utilisation du test.
1. Calcul du paramètre de l'échantillon mentionné dans la règle de décision
2. Application de la règle de décision

EX 1

Reprendre l'exercice 2 du chapitre '' Estimation ''
Au garage où sont stationnés les camions, le responsable affirme qu'il y a, en moyenne, 4 camions en panne par jour. Un des chauffeurs prétend qu'il y en a 6.
1. Construire un test permettant d'accepter ou non, au seuil de 5 % et au vu des résultats sur 30 jours, le point de vue du responsable.

Utiliser ce test avec l'échantillon de l'énoncée.
2. Même question pour la proposition du chauffeur.
3. Reprendre ces deux questions en utilisant le seuil de signification de 1 %.

EX 2

Reprendre l'exercice 3 du chapitre '' Estimation ''
1. Construire un test permettant d'accepter ou de rejeter , au seuil de signification de 5 %, l'hypothèse selon laquelle, au vu d'un sondage portant sur 1068 personnes interrogées, le candidat sera élu au premier tour.

Utiliser ce test avec l'échantillon de l'énoncée.
2. Même question au seuil de signification de 1 %.

COMPARAISON DES MOYENNES DE DEUX POPULATIONS.

Exemple:

Une entreprise reçoit un lot de 500 pièces d'une entreprise *A,* et un lot de 800 pièces d'une entreprise *B.* Dans chaque cas la masse moyenne indiquée par les fournisseurs est de 780 g.
Un échantillon est prélevé dans chaque lot, d'effectif 36 pour le lot *A*, d'effectif 50 pour le lot *B*. Les moyennes des échantillons sont: $\bar{x}_A = 774{,}7$ et $\bar{x}_B = 779{,}6$.

La question que l'on peut se poser est: la différence entre ces deux moyennes provient-elle d'une différence entre les productions des deux fournisseurs ou du choix des échantillons ?
Autrement dit, comment construire et utiliser un test permettant de décider, à partir des échantillons prélevés, s'il y a une différence significative, au seuil de 5 % par exemple entre les moyennes des masses des pièces livrées par les deux fournisseurs ?

Les distributions des moyennes $\overline{X}_A$ et $\overline{X}_B$ suivent respectivement les lois normales $N\left(m_A, \frac{\sigma_A}{\sqrt{n_A}}\right)$ et $N\left(m_B, \frac{\sigma_B}{\sqrt{n_B}}\right)$.

Supposons que ces deux variables aléatoires $\overline{X}_A$ et $\overline{X}_B$ sont indépendantes.
Par définition, la variable aléatoire D = $\overline{X}_B - \overline{X}_A$ associe à tout échantillon de taille 36 prélevé dans la population *A* et à tout échantillon de taille 50 prélevé dans la population *B* la différence des moyennes de l'échantillon *B* et de l'échantillon *A*.

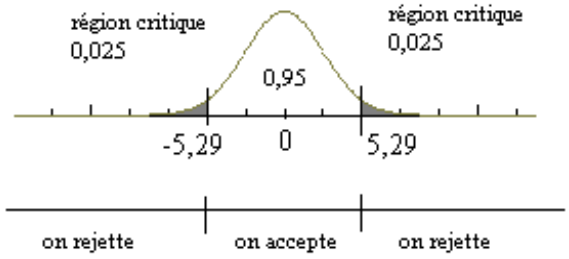
$D = \overline{X}_B - \overline{X}_A$ suit une loi normale et:

$E(D) = E(\overline{X}_B) - E(\overline{X}_A) = m_B - m_A$; $V(D) = V(\overline{X}_B) + V(\overline{X}_A) = \frac{\sigma_B^2}{n_B} + \frac{\sigma_A^2}{n_A}$

L'écart-type de *D* est donc: $\sigma_D = \sqrt{\frac{\sigma_B^2}{n_B} + \frac{\sigma_A^2}{n_A}} \approx 2{,}7$. *D* suit donc la loi normale $N(m_B - m_A ; 2{,}7)$.

Construction du test.

$H_0 : m_B = m_A$; $H_1 : m_B \neq m_A$
On teste donc la validité de l'hypothèse: la moyenne des masses des pièces sur l'ensemble de chaque livraison est la même pour les fournisseurs *A* et *B* , avec une probabilité de 0,95.
Sous H_0, *D* suit la loi normale $N(0; 2{,}7)$.
On obtient $p(-5{,}29 \leq D \leq 5{,}29) = 0{,}95$



Règle de décision.
On prélève un échantillon aléatoire de taille $n_A = 36$ dans la population *A* et on calcule sa moyenne $\overline{x}_A$; on fait de même avec la population *B*, avec $n_B = 50$. Soit $d = \overline{x}_B - \overline{x}_A$.

- Si $d \in [-5{,}29 ; 5{,}29]$ on accepte H_0.
- Si $d \notin [-5{,}29 ; 5{,}29]$ on rejette H_0.

Utilisation du test dans l'exemple donné.

$d = 779{,}6 - 774{,}7 = 4{,}9$ donc on accepte H_0.
Au seuil de 5 % il n'y a pas de différence significative entre les moyennes des masses des pièces livrées par les deux fournisseurs.

Autre problème:

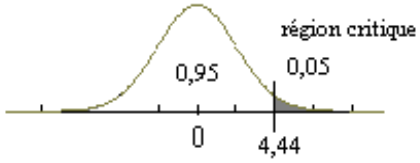
Comment construire et utiliser un test permettant de décider, à partir des mêmes échantillons, si la moyenne des masses des pièces livrées par le fournisseur *B* est significativement supérieure, au seuil de 5 %, à celle du fournisseur *A* ?
Il semble alors naturel de prendre pour hypothèse H_0: $m_B > m_A$, mais, ne connaissant pas la valeur numérique de $m_B - m_A$, il n'est pas possible d'obtenir des résultats numériques pour la variable aléatoire *D*.
On doit donc choisir une égalité pour H_0, dans tous les cas. Ici on choisit $m_B = m_A$.
$m_B > m_A$ est alors l'hypothèse alternative H_1 du test unilatéral.
H_1 est acceptée lorsque H_0 est rejetée et inversement.

Construction du test.

$H_0 : m_B = m_A$; $H_1 : m_B > m_A$

Détermination de la région critique au seuil de 5 %.

Sous H_0, *D* suit la loi normale N (0; 2,7).
On obtient p ($D \leq 4{,}44$) = 0,95



Règle de décision.

On prélève un échantillon aléatoire de taille $n_A = 36$ dans la population *A* et on calcule sa moyenne $\bar{x}_A$; on fait de même avec la population *B*, avec $n_B = 50$. Soit $d = \bar{x}_B - \bar{x}_A$.

- Si $d \leq 4{,}44$ on accepte H_0 et on rejette H_1.
- Si $d > 4{,}44$ on rejette H_0 et on accepte H_1.

Utilisation du test dans l'exemple donné.

$d = 779{,}6 - 774{,}7 = 4{,}9$ or $4{,}9 > 4{,}44$ donc on rejette H_0 et on accepte H_1.
Au seuil de 5 % la moyenne des masses des pièces livrées par le fournisseur *B* est significativement supérieure à celle du fournisseur *A*.

Remarques.

La définition de la procédure est capitale dans la prise de décision: au même seuil de 5 %, et à partir des mêmes échantillons, les conclusions ont été différentes avec un test bilatéral et un test unilatéral.
Dans le cas de test de validité d'hypothèse relatif à un pourcentage, la démarche est analogue en remplaçant la variable aléatoire $\bar{X}$ par la variable aléatoire F.

EX 3.

Une entreprise s'interroge sur l'efficacité d'un dispositif destiné à améliorer la qualité d'une fabrication.
Sans utiliser ce dispositif, un pourcentage p inconnu de pièces fabriquées sont correctes, les autres devant être retravaillées.
Sur un échantillon de 100 pièces, on observe que 25 sont correctes.
En utilisant ce dispositif, le pourcentage inconnu de pièces correctes est noté p'.
Sur un échantillon de 400 pièces, on observe que 136 sont correctes.
On suppose que tous les échantillons intervenant ici peuvent être considérés comme prélevés au hasard et avec remise et que la variable aléatoire F (respectivement F') qui, à un tel échantillon de taille $n = 100$ (respectivement $n' = 400$) associe son pourcentage de pièces correctes, suit une loi normale.
On suppose que F et F' sont indépendantes.
On prend pour valeur de l'écart-type de F ou de F' l'estimation ponctuelle fournie par l'échantillon correspondant.
Peut-on, au seuil de signification de 5 %, considérer que le dispositif améliore la qualité de la fabrication?

EX 4.

Une entreprise fabrique des pots de peinture.
Elle les fait livrer habituellement par lots de 20 pots ou de 100 pots. On se propose d'étudier les variations de la quantité d'un certain produit A contenu dans chaque pot.

Partie A

On suppose que la production totale de l'entreprise est très importante et que 7,5 % des pots fabriqués contiennent plus de 110 g de substance A. On note X la variable aléatoire qui à tout tirage aléatoire de 20 pots (tirage considéré comme tirage avec remise) associe le nombre de pots contenant plus de 110 g de substance A. On note de même Y la variable associée dans le cas de tirages de 100 pots.

1) Préciser la loi de X.
2) Calculer au millième le plus proche la probabilité de l'événement « X = 1 ».
3) Préciser la loi de Y.
4) On veut approcher la loi de Y par une loi de Poisson de même espérance mathématique. Préciser le paramètre de cette loi de Poisson.
5) En supposant que Y suive effectivement la loi de Poisson ainsi définie, donner une approximation au millième le plus proche de la probabilité de l'événement « $Y \leq 6$ ».

Partie B

On a contrôlé le dosage du produit A à la sortie de deux chaînes de fabrication.
Deux échantillons de 100 pots ont été analysés ; l'un provient de la chaîne l, l'autre de la chaîne 2.

Le tableau suivant donne la répartition de l'échantillon de la chaîne l en fonction de la masse de produit A exprimée en grammes.

m (en g)	[100,102[	[102,104[	[104,106[	[106,108[	[108,110[	[110,112[	[112,114[	[114,116[
Effectifs	1	3	25	32	27	6	4	2

On donne des valeurs approchées de la moyenne m_2 et de l'écart type σ_2 de l'échantillon fabriqué par la chaîne 2: $m_2 = 107$ et $\sigma_2 = 2$ (en grammes).

Dans les questions 1 et 2 les valeurs seront arrondies au dixième le plus proche.

1) En prenant les centres des classes, calculer une approximation de la moyenne m_1 et de l'écart type σ_1 de l'échantillon issu de la chaîne l.

2) En considérant les résultats obtenus dans la première question, donner les estimations ponctuelles :

a) des quantités moyennes μ_1 et μ_2 de produit A pour les productions de ces deux chaînes,

b) des écarts types s_1 et s_2 correspondants.

3) On se propose de savoir si la différence des moyennes observées dans les deux échantillons est due à des fluctuations d'échantillonnage ou si la chaîne de fabrication l produit des pots contenant davantage de produit A que la chaîne 2.
On note $\overline{X_1}$ la variable aléatoire qui a tout échantillon aléatoire de 100 pots provenant de la chaîne l associe la quantité moyenne de produit A dans cet échantillon.
On note $\overline{X_2}$ la variable aléatoire qui a tout échantillon aléatoire de 100 pots provenant de la chaîne 2 associe la quantité moyenne de produit A dans cet échantillon.

On admettra que:

- $\overline{X_1}$ suit une loi normale de paramètres μ_1 et $\frac{s_1}{10}$
- $\overline{X_2}$ suit une loi normale de paramètres μ_2 et $\frac{s_2}{10}$
- $\overline{X_1}$ et $\overline{X_2}$ sont des variables aléatoires indépendantes
- $D = \overline{X_1} - \overline{X_2}$ suit une loi normale.

On choisit l'hypothèse nulle H_0 : « $\mu_1 = \mu_2$ » contre l'hypothèse alternative H_1 : « $\mu_1 > \mu_2$ »

a) Calculer la variance de la variable aléatoire *D*. On appelle $\sigma(D)$ son écart type.
Vérifier que $\sigma(D) \approx 0{,}32$.
b) Calculer au centième le plus proche le réel a tel que $P(D < a) = 0{,}99$.
c) L'hypothèse nulle H_0 est-elle acceptée ou rejetée (au seuil de 1 %) ?