Chapitre 9

Statistiques inférentielles

9.1 Introduction – vocabulaire

Pour étudier une population statistique, on a recours à deux méthodes :

- la **méthode exhaustive** (ou *recensement*) : on examine chacun des éléments de la population. En général, cette méthode est jugée trop longue.
- la **méthode des sondages** : on n'examine qu'une partie de la population pour essayer d'en déduire des informations sur la totalité de la population. Cette méthode comprend deux parties :
  - l'**échantillonage** qui permet de passer de la population totale à une partie seulement de cette population (l'*échantillon*).
  - l'**estimation** qui permet d'induire, à partir des résultats observés sur l'échantillon, des informations sur la population totale.

Nous ne nous préoccuperons pas ici des problèmes concernant l'échantillonage. Notre propos sera seulement d'examiner deux méthodes différentes d'estimation.

9.2 Principe de la théorie

On considère une population P d'effectif N. On suppose que, pour le caractère observé, la moyenne de P est m alors que son écart-type est σ. Ce sont ces deux valeurs que nous voudrions retrouver à partir des échantillons. Supposons donc maintenant que nous disposons de k échantillons de P, chacun d'entre eux étant d'effectif n. On note $E_1, E_2, \ldots, E_k$ ces k échantillons, de moyennes respectives $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k$, et d'écart-type respectifs $\sigma_1, \sigma_2, \ldots \sigma_k$.
L'ensemble $\overline{X} = \{\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k\}$ est une série statistique d'effectif k, série que l'on appelle *distribution des moyennes*. La théorie montre alors que

$$E\big(\overline{X}\big) = m, \qquad \text{et} \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

De plus, pour $n > 30$, la variable aléatoire $T = \dfrac{\sqrt{n}}{\sigma}\left(\overline{X} - m\right)$ suit une loi normale $\mathscr{N}(0, 1)$.

9.3 Estimation ponctuelle

Connaissant la moyenne $\overline{x}$ et l'écart-type σ' d'un échantillon de taille n, il s'agit d'estimer la moyenne m de la population totale P. L'estimation ponctuelle est la méthode naïve qui consiste à confondre la moyenne de l'échantillon avec la moyenne de la population totale. On dira que $\overline{x}$ est une *estimation ponctuelle de la moyenne* m.
Par analogie on est tenté de choisir la variance σ'^2 d'un échantillon prélevé au hasard comme estimation ponctuelle de la variance inconnue σ^2 d'une population. Mais, en procédant ainsi on à tendance à sous-estimer la variance de la population.
On choisit le nombre $\dfrac{n}{n-1}\sigma' 2$, comme estimation ponctuelle de la variance inconnue σ^2 de cette population.
On choisit le nombre $\sqrt{\dfrac{n}{n-1}}\sigma'$, comme estimation ponctuelle de l'ecart-type inconnue σ de cette population.

9.4 Estimation par intervalle de confiance

On considère une population P d'effectif N. On suppose que, pour le caractère observé, la moyenne, inconnue, de P est m alors que son écart-type, connu, est σ. On prélève au hasard, et avec remise, une succession d'échantillons de même effectif n dont on calcule les moyennes respectives : $\overline{x}_1$ pour le premier, $\overline{x}_2$ pour le deuxième, et ainsi de suite.
Notons maintenant $\overline{X}$ la variable aléatoire qui associe à un échantillon E_i sa moyenne x_i. La variable $\overline{X}$ prend donc successivement les valeurs $\overline{x}_1$, $\overline{x}_2$,...
Pour finir, on suppose également que les conditions sont réunies pour pouvoir utiliser une conséquence du théorème de la limite centrée et faire l'approximation que X suit la loi normale $\mathcal{N}(m, \sigma/\sqrt{n})$. Autrement dit que la variable aléatoire $T = \sqrt{\frac{n}{\sigma}}(\overline{X} - m)$ suit la loi normale $\mathcal{N}(0,1)$. On aura alors, pour tout $t \geq 0$,

$$P(-t \leq T \leq t) = 2\Pi(t) - 1$$

9.4.1 Calcul sur un exemple : intervalle de confiance à 95%

Par exemple, si on veut obtenir un intervalle ayant 95% de chances de contenir la moyenne m de la population P, on procède de la manière suivante :
• On a $2\Pi(t) - 1 = 0,95 \iff \Pi(t) = 0,975$. Avec la table donnée dans le formulaire, on voit que cette valeur est obtenue pour $t = 1,96$. On a donc

$$\begin{aligned} & P\left(-1,96 \leq \frac{\sqrt{n}}{\sigma}(\overline{X} - m) \leq 1,96\right) = 0,95 \\ \iff \quad & P\left(-1,96\frac{\sigma}{\sqrt{n}} \leq (\overline{X} - m) \leq 1,96\frac{\sigma}{\sqrt{n}}\right) = 0,95 \\ \iff \quad & P\left(m - 1,96\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq m + 1,96\frac{\sigma}{n}\right) = 0,95 \end{aligned}$$

Autrement dit, **avant de prélever un échantillon** de taille n dans la population, il y a 95% de chances pour que cet échantillon ait une moyenne entre

$$m - 1,96\frac{\sigma}{n} \qquad \text{et} \qquad m + 1,96\frac{\sigma}{n}$$

• Comme m est inconnu, on se sert des résultats précédents pour encadrer m :

$$\begin{aligned} & P\left(-\overline{X} - 1,96\frac{\sigma}{\sqrt{n}} \leq -m \leq -\overline{X} + 1,96\frac{\sigma}{\sqrt{n}}\right) = 0,95 \\ \iff \quad & P\left(\overline{X} + 1,96\frac{\sigma}{\sqrt{n}} \geq m \geq \overline{X} - 1,96\frac{\sigma}{\sqrt{n}}\right) = 0,95 \\ \iff \quad & P\left(\overline{X} - 1,96\frac{\sigma}{\sqrt{n}} \leq m \leq \overline{X} + 1,96\frac{\sigma}{\sqrt{n}}\right) = 0,95 \end{aligned}$$

Ainsi, **avant de prélever un échantillon** de taille n dans la population, il y a 95% de chances pour la moyenne $\overline{x}$ de cet échantillon vérifie

$$\overline{x} - 1,96\frac{\sigma}{\sqrt{n}} \leq m \leq \overline{x} + 1,96\frac{\sigma}{\sqrt{n}}$$

En revanche, **après** le prélèvement, il n'y a plus de probabilité à envisager : il est vrai ou faux que la moyenne m se situe dans l'intervalle envisagé $\left[\overline{x} - 1,96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1,96\frac{\sigma}{\sqrt{n}}\right]$. Cet intervalle est appelé *intervalle de confiance de la moyenne de la population avec le coefficient de confiance* 95% (ou *avec le risque* 5%).

9.4.2 Cas général

On fonctionne exactement sur le même principe : un coefficient de confiance choisi à l'avance permet de définir un nombre positif t tel que $P(-t \leq T \leq t) = 2\Pi(t) - 1$ soit égal à ce coefficient de confiance.
Par exemple, $2\Pi(t) - 1 = 0,99$ si et seulement si $\Pi(t) = 0,995$, ce qui correspond à $t = 2,58$ (d'après la table de la loi normale $\mathcal{N}(0,1)$).
En reprenant tous les calculs ci-dessus, oin obtient alors le résultat suivant :

Théorème 1 *L'intervalle* $\left[\overline{x} - t\frac{\sigma}{\sqrt{n}}, \overline{x} + t\frac{\sigma}{\sqrt{n}}\right]$ *est l'intervalle de confiance de la moyenne* m *de la population avec le coefficient de confiance* $2\Pi(t) - 1$*, ayant pour centre la moyenne* $\overline{x}$ *de l'échantillon considéré.*

Dans la pratique, on utilise souvent des coefficients de confiance de 95%, ce qui correspond à $t = 1,96$, ou à 99%, ce qui correspond à $t = 2,58$.

9.5 Test de validité d'hypothèse :

9.5.1 nature du problème

Depuis quelques décennies, on assiste à une « entrée en force » des méthodes statistiques dans le domaine réglementaire, lequel conduit à la **prise de décision** : on a ou on n'a pas le droit de ...
En particulier, l'augmentation des échanges commerciaux et des liens économiques entre les pays s'accompagne d'accords destinés à fixer les règles communes. Les statistiques inférentielles trouvent là un immense champ d'applications.
Cela se traduit par des règlementations définissant dans chaque cas particulier une procédure destinée à préciser sans ambiguïté :

- comment un ou plusieurs échantillons doivent être prélevés dans la population étudiée ;
- quelles mesures doivent être effectuées sur ce ou ces échantillon(s) ;
- quelle décision doit être prise à propos de l'ensemble de la population.

Une telle procédure s'appelle, en statistique, un *test de validité d'hypothèse*.
De manière plus générale, il s'agit, à partir de l'étude d'un ou plusieurs échantillons, de prendre des décisions concernant l'ensemble de la population.

9.5.2 Exemple de test : comparaison de la moyenne d'une population à un nombre fixé

Une société s'approvisionne en pièces brutes qui, conformément aux conditions fixées par le fournisseur, doivent avoir une masse moyenne de 780 grammes.Au moment où 500 pièces sont réceptionnées, on en prélève au hasard un échantillon de 36 pièces dont on mesure la masse. On obtient les résultats suivants :

Masses des pièces (en grammes)	Nombre de pièces
[745, 755 [	2
[755, 765 [	6
[765, 775 [	10
[775, 785 [	11
[785, 795 [	5
[795, 805 [	2

La masse moyenne des pièces de l'échantillon est de $774,7$ g.
En supposant que l'écart-type des masses pour la population des 500 pièces est $\sigma = 12,5$ g, on obtient $[770,61\,;778,79]$ comme intervalle de confiance à 95% de la moyenne inconnue m de cette population.
Présentation du problème :
Peut-on considérer que les 500 pièces de la population ont une masse moyenne de 780 g, comme le prévoient les conditions fixées par le fournisseur ? Autrement dit, doit-on ou non accepter la livraison de ces 500 pièces au vu du résultat obtenu sur l'échantillon ?

Hypothèse nulle

On suppose que la moyenne de la population est 780. C'est *l'hypothèse nulle*, notée H_0 : $m = 780$. Alors, la variable aléatoire $\overline{X}$ qui, à tout échantillon aléatoire non exhaustif de taille $n = 36$, associe la moyenne de cet échantillon suit approximativement la loi normale $\mathcal{N}(m, \sigma/\sqrt{n})$ où $m = 780$.
Donc $\frac{\sqrt{n}}{\sigma}\left(\overline{X} - m\right)$ suit la loi normale centrée réduite $\mathcal{N}(0,1)$. Donc, pour tout $t > 0$, on a

$$P\left(-t \leq \frac{\sqrt{n}}{\sigma}\left(\overline{X} - m\right) \leq t\right) = 2\Pi(t) - 1$$

où la valeur de $\Pi(t)$ est lue dans la table de la loi normale $\mathcal{N}(0,1)$ du formulaire de mathématiques. Donc, pour tout $t > 0$,

$$P\left(m - \frac{t\sigma}{\sqrt{n}} \leq \overline{X} \leq m + \frac{t\sigma}{\sqrt{n}}\right) = 2\Pi(t) - 1$$

En particulier, si $t = 1,96$, on a $2\Pi(t) - 1 = 0,95$, d'où, ici,

$$P\big(775,92 \leq \overline{X} \leq 784,08\big) = 0,95.$$

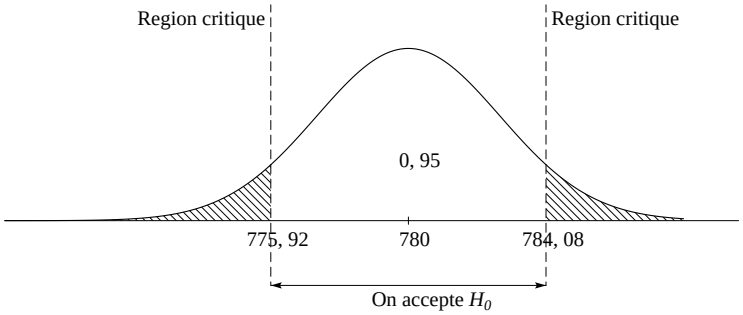
Ainsi, en supposant que $m = 780$, on sait, avant de prélever un échantillon aléatoire de taille 36, que l'on a 95% de chance que sa moyenne soit dans l'intervalle $[775,92\,;784,08]$. Autrement dit, si H_0 est vraie, il n'y a que 5% de chance de prélever un échantillon aléatoire de taille 36 dont la moyenne soit inférieure à $775,92$ ou supérieure à $784,08$.

Règle de décision, région critique

On fixe alors la règle de décision suivante : on prélève un échantillon aléatoire non exhaustif de taille $n = 36$ et on calcule sa moyenne $\bar{x}$.

Si $\bar{x} \in [775,92\,;784,08]$, on accepte H_0

Si $\bar{x} \notin [775,92\,;784,08]$, on rejette H_0



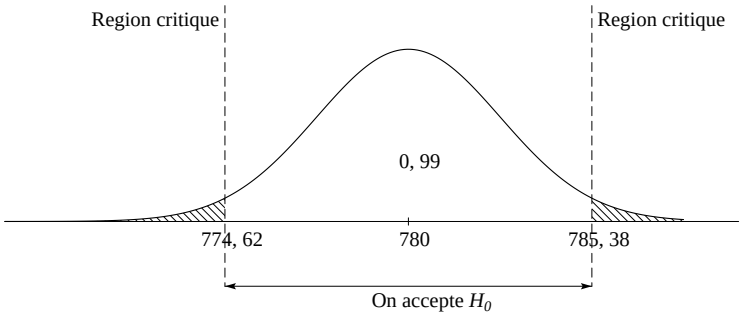
Si H_0 est vraie, on prend donc le risque de se tromper dans 5% des cas en rejetant à tord H_0. On définit ainsi une *région critique au seuil* $\alpha = 5\%$. Le seuil α est la probabilité de rejeter H_0 alors que H_0 est vraie. Il correspond à l'*erreur de première espèce*. En général, on fixe *a priori* la valeur de α (ici égal à $0,05$). Dans l'exemple qui nous occupe, on a $\bar{x} = 774,7$ pour l'échantillon considéré. On a $\bar{x} < 775,92$ et on rejette l'hypothèse H_0. Au seuil de 5%, on considère que les 500 pièces de la population n'ont pas une moyenne de 780 g et on refuse la livraison.

Erreur de seconde espèce

On aurait pu choisir un seuil de 1% pour diminuer le risque de rejeter H_0 alors que H_0 est vraie. On a

$$p(774,62 \leq \overline{X} \leq 785,38) = 0,99.$$

Au seuil de 1%, on accepte H_0 puisque $\bar{x}$ appartient à l'intervalle considéré, et on accepte alors la livraison des 500 pièces.



Mais, en acceptant H_0 au seuil de 1%, on court un second risque : celui d'accepter H_0 alors que H_0 est fausse : c'est l'*erreur de seconde espèce*, dont la probabilité est notée β. En général, lorque la taille n de l'échantillon est fixée, on a α qui diminue lorsque β augmente, et réciproquement. Le seule façon de diminuer en même temps α et β 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.

Hypothèse alternative

Il faut définir plus précisément le cas où H_0 est fausse. Dans ce qui précède, on a choisi implicitement $m \neq 780$ comme *hypothèse alternative* H_1. Le test est alors *bilatéral*, car la région critique est située des deux côtés de la région où on accepte H_0. Si on décide, par exemple, de prendre $m < 780$ comme hypothèse alternative H_1, le test est alors *unilatéral* et la région critique est située entièrement d'un côté de la région où on accepte H_0.

Résumé

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

1. **Construction du test**
   (a) Choix de l'hypothèse nulle H_0 et de l'hypothèse alternative H_1.
   (b) Détermination de la région critique à un seuil α donné.
   (c) Enoncé de la règle de décision : si un paramètre du ou des échantillon(s) est dans la région critique, on rejette H_0, sinon on l'accepte.
2. **Utilisation du test**
   (a) Calcul du paramètre de l'échantillon mentionné dans la règle de décision,
   (b) Application de la règle de décision.

9.5.3 Exemple de test : comparaison de la moyenne de deux populations

Présentation du problème

Un second fournisseur B livre 800 pièces du même modèle. On prélève au hasard et avec remise un échantillon de 50 pièces dont on mesure la masse. On obtient les résultats suivants :

Masses des pièces(en grammes)	Nombre de pièces
$[745, 755[$	6
$[755, 765[$	12
$[765, 775[$	16
$[775, 785[$	11
$[785, 795[$	4
$[795, 805[$	1

La masse moyenne des pièces de l'échantillon est de $779,6$ alors que l'échantillon de 36 pièces provenant du premier fournisseur a pour moyenne $774,7$ g.
La différence de $4,9$ entre ces 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 ci-dessus, s'il y a une différence significative, au seuil de 5%, entre les moyennes des masses des pièces livrées par les deux fournisseurs ?

Un peu de théorie

Nous sommes en présence de deux échantillons extraits de deux populations correspondant aux deux fournisseurs A et B. On peut schématiser la situation de la façon suivante :

m_A : inconnue $\sigma_A = 12,5$	$n_A = 36$ $\overline{x}_A = 774,7$ $\sigma'_A = 12,36$ Echantillon

Population A

m_B : inconnue $\sigma_B = 12,1$	$n_B = 50$ $\overline{x}_B = 779,6$ $\sigma'_B = 11,99$ Echantillon

Population B

Soit $\overline{X}_A$ (resp $\overline{X}_B$) la variable aléatoire qui, à tout échantillon de taille $n_A = 36$ (resp $n_B = 50$) prélevé aléatoirement et avec remise dans la population A (resp B), associe la moyenne des masses de pièces de l'échantillon.
On se place dans le cas où $\overline{X}_A$ suit approximativement la loi normale $\mathcal{N}(m_A, \sigma_A/\sqrt{n_A})$ et $\overline{X}_A$ suit approximativement la loi normale $\mathcal{N}(m_B, \sigma_B/\sqrt{n_B})$.
Par définition, la variable aléatoire $D = \overline{X}_B - \overline{X}_B$ associe à tout échantillon de taille 36 ainsi prélevé dans la population A et à tout échantillon ainsi prélevé dans la population B la différence des moyennes de l'échantillon B et de l'échantillon A.
On suppose que les variables $\overline{X}_A$ et $\overline{X}_B$ sont **indépendantes**.
Alors $D = \overline{X}_B - \overline{X}_A$ suit une loi normale et

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

L'écart-type de D est donc $\sqrt{\dfrac{\sigma_B^2}{n_B} + \dfrac{\sigma_A^2}{n_A}} \approx 2,7$, et D suit une loi normale $\mathcal{N}(m_B - m_A\,;2,7)$.

Construction du test

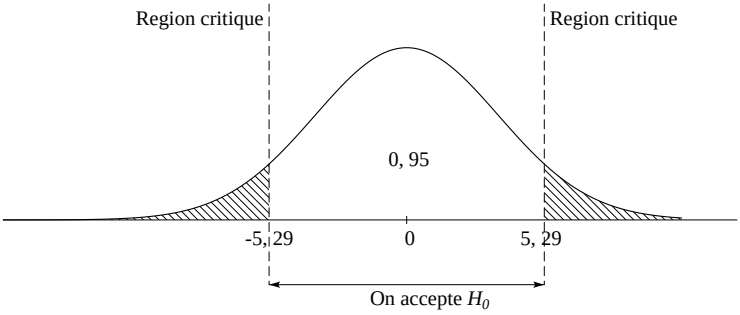
- *Choix de H_0 : $m_A = m_B$.*

  Choix de H_1 : $m_A \neq m_B$.

  Nous allons tester 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* ».

- *détermination de la région critique au seuil de* 5%

  Sous l'hypothèse H_0, D suit la loi normale $\mathcal{N}(0\,;2,7)$, donc $D/2,7$ suit la loi normale centrée réduite $\mathcal{N}(0,1)$. En particulier, on a $p(-t \leq D/2,7 \leq t) = 0,95$ lorsque $t = 1,96$, et donc $p(-5,29 \leq D \leq 5,29) = 0,95$.



- *Enoncé de la règle de décision*

  On prélève avec remise un échantillon aléatoire de taille $n_A = 30$ de la population A et on calcule sa moyenne $\bar{x}_A$; on fait de même pour la population B avec $n_B = 50$. On pose $d = \bar{x}_B - \bar{x}_A$.

si $d \in [-5,29\,;5,29]$ on accepte H_0.

si $d \notin [-5,29\,;5,29]$ on rejette H_0 et on accepte H_1.

Utilisation du test

- *Calcul de d*

On a $d = \bar{x}_B - \bar{x}_A = 779,6 - 774,7 = 4,9$

- *Application de la règle de décision*

Comme $4,9 \in [-5,29\,;5,29]$, on accepte H_0 : au seuil de 5%, il n'existe pas de différence significative entre les moyennes des masses des pièces livrées par les deux fournisseurs.