Dernière Leçon

Statistiques Inférentielles

Ce chapitre vous introduit dans le monde des *Statistiques inférentielles* (**Inférer** : tirer une conséquence de quelque proposition, de quelque fait, etc.) auquel vous êtes en fait constamment confronté(e) en tant que citoyen(ne) d'une société hautement médiatisée.

La statistique inférentielle est en effet la science qui permet de « *modéliser une partie observable du réel comme résultant d'un phénomène aléatoire pour lequel on envisage non pas une mais toute une famille de lois de probabilités possibles* » (J.P. Raoult - dossier APMEP statistique inférentielle - Déc 2005)

C'est un sujet très intéressant, exigeant une réflexion approfondie en classe sur les notions abordées, les conclusions à émettre, entraînant un débat intéressant, permettant d'avoir une démarche scientifique partant d'une expérimentation.

Malheureusement, nous sommes loin de disposer du temps nécessaire. Nous nous contenterons donc d'une préparation pure et simple aux exercices d'examen qui sont tous identiques, mis à part le contexte expérimental proposé.

I - Un exemple pour découvrir

Une population est constituée de 5 étudiants de BTS. Leur professeur de mathématiques adoré et vénéré veut estimer le temps moyen hebdomadaire consacré au travail personnel.

Étudiant	Temps de travail (secondes)
Albert	7
Bernard	3
Charles-Henri	6
Dolf	10
Évariste	4

Calculez moyenne et écart-type : $\mu =$, $\sigma =$.

Le problème, c'est qu'on n'a pas souvent accès à toute la population et on se contente d'échantillons.

Voyons ce qui se passe en prenant des échantillons de taille 3 de la population précédente.

Tout d'abord, combien y a-t-il d'échantillons de taille 3 ?

Remplissez le tableau suivant :

Échantillons	Données	Moyennes échantillonnales	Écarts-type échantillonaux

Calculez la moyenne des moyennes échantillonnales : $\mu_{\bar{x}} =$
Pour les écart-types : $\sigma_{\bar{x}} =$
Et alors ? En fait, en réalité, on n'a pas souvent accès aux données concernant la population entière : on se contente d'étudier un échantillon (sondage...) On voudrait alors savoir dans quelle mesure les résultats relevés sur un échantillon peuvent nous permettre d'*inférer* les résultats de toute la population.
Il existe un beau théorème en probabilités, le *Théorème de la Limite Centrale*, que le Russe Lyapounov a démontré en 1900. Il peut être appliqué dans le cas particulier de notre échantillonage. Il confirme que $\mu_{\overline{x}} = \mu$, il affirme que si la taille de l'échantillon est suffisamment grand ($n \geqslant 30$), alors les moyennes des échantillons suivent une distribution qui se rapproche d'une distribution normale $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$ et que $\sigma_{\overline{x}} \approx \frac{\sigma}{\sqrt{n}}$ pour de grandes valeurs de n.
Le problème, c'est qu'on ne connaît ni μ, ni $\mu_{\overline{x}}$ et idem pour les écarts-type. Il va donc falloir faire des *estimations*.

II - Estimation ponctuelle

On ne considère qu'une seule mesure. On ne cherche pas à calculer le risque d'erreur.
Dans ce cas

$\mu_{\overline{x}}$ est une estimation ponctuelle de la moyenne

et on admettra que

$\sigma_{\overline{x}}\sqrt{\frac{n}{n-1}}$ est une estimation ponctuelle de l'écart-type σ

III - Estimation de la moyenne par intervalle de confiance

On va cette fois essayer de déterminer un intervalle qui pourrait contenir la véritable valeur de μ avec un risque d'erreur décédé à l'avance.
Pour des échantillons de taille assez grande, la distribution des moyennes des échantillons suit la loi normale $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$.
On introduit comme d'habitude la variable $\mathrm{T} = \dfrac{\overline{\mathrm{X}} - \mu}{\frac{\sigma}{\sqrt{n}}}$.
On fixe à l'avance une probabilité α que T n'appartienne pas à un intervalle inconnu $[-t; t]$. Alors on a

$$1 - \mathrm{P}(-t \leqslant \mathrm{T} \leqslant t) = \alpha$$

c'est-à-dire :

$$1 - \Big(2\Pi(t) - 1\Big) = \alpha$$

ou encore

$$\Pi(t) = \frac{2-\alpha}{2}$$

On cherche t dans notre table et nous saurons encadrer μ.
En général, α vaut 1%, 5% ou 10%.
On a alors :

α	1%	5%	10%
t			

Un exemple !

À partir de la moyenne $\overline{x}$ d'un échantillon, on veut déterminer un intervalle qui contient la vraie valeur de la moyenne avec 5% de chance de se tromper.
Pour $\alpha = 0,05$, on a $t = 1,96$.
On a donc

$$P\left(-1,96 \leqslant \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leqslant 1,96\right) = 0,95$$

et après calculs :

$$P\left(\overline{x} - 1,96\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \overline{x} + 1,96\frac{\sigma}{\sqrt{n}}\right) = 0,95$$

IV - Estimation de la fréquence par intervalle de confiance

On peut montrer comme pour les moyennes qu'un intervalle de confiance de la fréquence p au seuil de α est :

$$\left[f - t\sqrt{\frac{f(1-f)}{n-1}}, f + t\sqrt{\frac{f(1-f)}{n-1}}\right]$$

Un exemple

Un sondage dans un lycée relève que sur les 500 personnes interrogées, 42% *sont mécontentes du menu de la cantine. Déterminez, au seuil de risque de* 1%*, un intervalle de confiance du pourcentage p de personnes mécontentes dans le lycée.*

$f = 0,42$, $n = 500$, $\alpha = 0,01$ donc $t =$
Un intervalle de confiance de p à 1% est :

V - Test de validité d'hypothèse

a. De quoi s'agit-il ?

Depuis quelques décennies, on assiste à une « entrée en force » des méthodes statistiques dans le domaine règlementaire, 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 doîvent être prélevés dans la population étudiée ;
– quelles mesures doîvent être effectuées sur ce ou ces échantillon(s) ;

– quelle décision doît 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.

b. Comparaison d'une moyenne à un nombre fixé : un exemple

Le problème : une moyenne est fixée et représente la norme. On prélève un échantillon d'une population censée suivre cette norme : doit-on, oui ou non, accepter la livraison ?

Examinons le problème suivant (BTS 2001) : un client réceptionne une commande. Il prélève un échantillon de 125 billes choisies au hasard et avec remise dans le lot reçu et constate que le diamètre moyen est égal à 25,1. On rappelle que pour les billes fabriquées par l'entreprise, la variable aléatoire X qui prend pour valeurs leurs diamètres suit une loi normale d'écart type 0,44. L'entreprise s'est engagée à ce que la moyenne des diamètres des billes fournies soit de 25. Le client décide de construire un test permettant de vérifier l'hypothèse selon laquelle le diamètre des billes du lot reçu est de 25.

Hypothèse nulle

On note H_0 et on appelle *hypothèse nulle* l'affirmation : $\mu = 25$.
Alors, la variable aléatoire $\overline{X}$ qui, à tout échantillon aléatoire de taille $n = 125$ associe la moyenne de cet échantillon suit approximativement la loi normale $\mathcal{N}\left(25; \frac{\sigma}{\sqrt{n}}\right)$, c'est-à-dire $\mathcal{N}(25; 0,039)$.
Cherchons le réel positif a vérifiant :

$$p(25 - a \leqslant \overline{X} \leqslant 25 + a) = 0,95$$

Introduisons $T = \frac{\overline{X}-25}{0,039}$. On cherche donc a tel que :

$$p(-\frac{a}{0,039} \leqslant T \leqslant \frac{a}{0,039}) = 0,95$$

$$2\Pi\left(\frac{a}{0,039}\right) - 1 = 0,95$$

$$\Pi\left(\frac{a}{0,039}\right) = 0,975$$

On obtient, à l'aide de la table :

$$a = 0,039 \times 1,96 \approx 0,08$$

Ainsi :

$$p\left(24,92 \leqslant \overline{X} \leqslant 25,08\right) = 0,95$$

En supposant $\mu = 25$, on sait, avant de prélever un échantillon aléatoire de taille 125, que l'on a 95% de chance que sa moyenne soit dans l'intervalle $[24,92; 25,08]$.
Autrement dit, dans ces conditions, il n'y a que 5% de chances d'obtenir une moyenne en dehors de cet intervalle.

Règle de décision

– Si la moyenne de l'échantillon appartient à $[24,92; 25,08]$, on accepte H_0 ;
– sinon, on rejette H_0
Le seuil de 5% est la probabilité de rejeter H_0 alors que H_0 est vraie.
Ici, on a mesuré une moyenne de 25,1 sur l''échantillon. On a $25,1 \notin [24,92; 25,08]$. On rejette donc l'hypothèse H_0. Au seuil de 5%, on considère que le stock entier de billes n'a pas la moyenne annoncée par le fabricant de 25 et on refuse la livraison.

Hypothèse alternative

Dans l'exemple précédent, on a choisi comme *hypothèse alternative* H_1 le cas où $\mu \neq 25$. On dit alors que le test est *bilatéral* car il existe deux régions critiques, à l'extérieur de l'intervalle $[24,92; 25,08]$.
On peut plutôt préférer un test *unilatéral* (ce sera précisé dans l'énoncé...). Par exemple, on peut prendre $\mu > 25,08$ comme hypothèse alternative H_1. La région critique est alors située d'un seul côté de la région où on accepte H_0.

VI - Des exercices

Exercice 1

Une entreprise fabrique en grande quantité des tiges en plastique de longueur théorique 100 mm. Un client reçoit un lot important de tiges de ce type. Il veut vérifier que la moyenne μ de l'ensemble des longueurs, en mm, des tiges constituant ce lot est égale à la longueur théorique.
On note L la variable aléatoire qui, à chaque tige prélevée au hasard dans le lot, associe sa longueur en mm. La variable aléatoire L suit la loi normale de moyenne inconnue μ et d'écart-type 0,16.
On désigne par $\overline{\text{L}}$ la variable aléatoire qui, à chaque échantillon aléatoire de 90 tiges prélevé dans le lot, associe la moyenne des longueurs de ces tiges (le lot est assez important pour que l'on puisse assimiler ces prélèvements à des tirages avec remise). $\overline{\text{L}}$ suit la loi normale de moyenne μ et d'écart-type $\sigma = \dfrac{0,16}{\sqrt{90}} \approx 0,017$.

Le client construit un test d'hypothèse :
- L'hypothèse nulle est $\mathbf{H}_0$: $\mu = 100$.
- L'hypothèse alternative est $\mathbf{H}_1$: $\mu \neq 100$.
- Le seuil de signification est fixé à 5 %.

1. Sous l'hypothèse nulle $\mathbf{H}_0$ déterminer le réel positif h tel que :

$$\text{P}(100 - h < \overline{\text{L}} < 100 + h) = 0,95.$$

2. Énoncer la règle de décision permettant d'utiliser ce test.

3. Le client prélève un échantillon aléatoire de 90 tiges dans la livraison et il constate que la moyenne des longueurs de l'échantillon est de 100,04 mm. Le client estime que le fournisseur n'a pas respecté ses engagements et renvoie tout le lot.
Le client a-t-il raison ? Justifier votre réponse.

Exercice 2

Une entreprise spécialisée produit des boules de forme sphérique en grande série.
Le responsable de la qualité cherche à analyser la production. Il mesure pour cela le diamètre des boules d'un échantillon (E) de 50 pièces. La moyenne obtenue sur l'échantillon (E) amène à se poser la question : « Le diamètre moyen m des boules fabriquées est-il strictement inférieur à 73 mm ? ».
Pour cela, on construit un test d'hypothèse au risque de 5 %.
L'hypothèse nulle H_0 est : $m = 73$;
L'hypothèse alternative H_1 est : $m < 73$.
On admet que la variable aléatoire $\overline{\text{D}}$, qui mesure le diamètre moyen sur un échantillon de 50 boules prélevées au hasard et avec remise, suit une loi normale de moyenne 73 et d'écart type $\dfrac{0,2}{\sqrt{50}}$.

1. Calculer le nombre réel a tel que $p(\overline{\text{D}} \geqslant 73 - a) = 0,95$.

2. Énoncer la règle de décision du test.

3. Au risque de 5 % et au vu de l'échantillon (E), que peut-on conclure ?

Exercice 3

Une entreprise fabrique, en grande quantité, des tiges métalliques cylindriques pour l'industrie. Leur longueur et leur diamètre sont exprimés en millimètres.
Dans cette question on s'intéresse au diamètre des tiges, exprimé en millimètres.
On prélève au hasard et avec remise un échantillon de 50 tiges dans la production d'une journée.
Soit $\overline{\text{D}}$ la variable aléatoire qui, à tout échantillon de 50 tiges prélevées au hasard et avec remise dans la production d'une journée, associe la moyenne des diamètres des tiges de cet échantillon.
On suppose que $\overline{\text{D}}$ suit la loi normale de moyenne inconnue μ et d'écart type $\dfrac{\sigma}{\sqrt{50}}$ avec $\sigma = 0,19$.
Pour l'échantillon prélevé, la moyenne obtenue, arrondie à 10^{-2} est $\overline{x} = 9,99$.

1. À partir des informations portant sur cet échantillon, donner une estimation ponctuelle de la moyenne μ des diamètres des tiges produites dans cette journée.

2. Déterminer un intervalle de confiance centré sur $\overline{x}$ de la moyenne μ des diamètres des tiges produites pendant la journée considérée, avec le coefficient de confiance 95 %.
3. On considère l'affirmation suivante : « la moyenne μ est obligatoirement dans l'intervalle de confiance obtenu à la question 2 ».
   Est-elle vraie ? (On ne demande pas de justification).

Exercice 4

On considère une grande quantité de pièces devant être livrées à une chaîne d'hypermarchés. On considère un échantillon de 100 pièces prélevées au hasard dans cette livraison. La livraison est assez importante pour que l'on puisse assimiler ce tirage à un tirage avec remise.
On constate que 96 pièces sont sans défaut.

1. Donner une estimation ponctuelle de la fréquence inconnue p des pièces de cette livraison qui sont sans aucun défaut.
2. Soit F la variable aléatoire qui, à tout échantillon de 100 pièces prélevées au hasard et avec remise dans cette livraison, associe la fréquence des pièces de cet échantillon qui sont sans défaut.

   On suppose que F suit la loi normale de moyenne p et d'écart type $\sqrt{\dfrac{p(1-p)}{100}}$, où p est la fréquence inconnue des pièces de la livraison qui sont sans aucun défaut.
   Déterminer un intervalle de confiance de la fréquence p avec le coefficient de confiance de 95%.

Exercice 5

La machine se déréglant dans le temps, on veut tester la moyenne m des longueurs des barres produites par la machine. On se demande si on peut accepter, au seuil de risque de 5 %, l'hypothèse selon laquelle la moyenne m des longueurs des barres est encore de $92,50$ cm.
Pour cela, on construit un test d'hypothèse bilatéral.
On suppose que la variable aléatoire X, qui à tout échantillon de 30 barres de métal prélevées au hasard associe la moyenne des longueurs en centimètres des barres de l'échantillon, suit une loi normale de moyenne m et d'écart type $0,03$.
On choisit l'hypothèse nulle H_0 : ´$m = 92,50$ˇ.

1. Donner l'hypothèse alternative H_1.
2. Sous l'hypothèse H_0, calculer le réel h tel que $P(92,5 - h < X < 92,5 + h) = 0,95$.
3. Énoncer la règle de décision du test.
4. On prélève un échantillon de 30 barres extraites au hasard dans la production de la machine, on obtient les résultats suivants :

Longueurs (en cm)	92,1	92,2	92,3	92,4	92,5	92,6	92,7	92,8	92,9
Nombre de barres	3	2	6	5	5	3	2	2	2

   Au vu des résultats de cet échantillon, peut-on admettre au seuil de risque de 5 %, l'hypothèse selon laquelle la moyenne m des longueurs des barres est encore de 92,50 cm ?