STATISTIQUE INFERENTIELLE

TRAVAUX DIRIGES

EXERCICE 1:

Une population est composée de 3 salariés A, B et C âgés respectivement de 23, 37et 45 ans.

1. On choisit au hasard un salarié.
   Définir ε, Ω, la probabilité P et la variable aléatoire X étudiée.
2. Calculer E(X) = μ et Var(X) = σ^2 . Que représente E(X) ? Que représente la Var(X) ?
3. On choisit au hasard un échantillon de 2 salariés.
   Définir ε_n, E_n, la probabilité P^* et les variables aléatoires X_1 et $\bar{X}_n$.
4. Calculer E($\bar{X}_n$) et Var($\bar{X}_n$). Retrouvez les formules du cours.

EXERCICE 2:

Une population est composée de 3 individus A, B et C dont les résultats de vote pour un certain candidat sont respectivement les suivants NON, NON et OUI.

1. On choisit au hasard un individu.
   Définir ε, Ω, la probabilité P et la variable aléatoire X étudiée.
2. Calculer E(X) et Var(X). Que représente E(X) ?
3. On choisit au hasard un échantillon de 2 individus.
   Définir ε_n, E_n, la probabilité P^* et la variable aléatoire P_n.
4. Calculer E(P_n) et Var(P_n). Retrouvez les formules du cours.

EXERCICE 3 :

Le poids de paquets de poudre de lessive, à l'issue de l'empaquetage, est supposé suivre une loi normale $N(\mu, \sigma)$ dont l'écart-type σ est supposé connu et égal à 5 g. σ représente la variabilité du poids due à l'imprécision de la machine. Le poids marqué sur les paquets est de 710g.
Toutes les heures, 10 paquets sont prélevés au hasard et pesés.
On obtient pour une heure donnée, pour un échantillon de 10 paquets un poids moyen de 707g.

1. Donner un estimateur puis une estimation du poids moyen des paquets de lessive.
2. Donner un intervalle de confiance à 90% , puis à 95% pour le poids moyen des paquets de lessive
3. Déterminer α (à l'unité près) pour qu'au seuil de risque α % un intervalle de confiance du poids moyen des paquets de lessive soit [705g;709g]

EXERCICE 4 :

Une firme nationale de sondages d'opinion a effectué pour le compte d'une compagnie d'assurance, une étude sur les besoins financiers et la satisfaction des clients. Dans la section du questionnaire concernant les fonds communs de placement, on demande aux clients de

donner la valeur (en euros) de tous les fonds communs de placement qu'ils possèdent. Voici les résultats pour un échantillon aléatoire de 20 clients :
Fond commun de placement
93850 121500 166675 173000 81580
172450 80515 191000 105630 192100
151975 148000 173400 138330 142500
149660 120225 149375 131170 85600
On suppose que la valeur actuelle des fonds communs de placement est distribuée normalement.

1. Donner un estimateur puis une estimation ponctuelle de la valeur moyenne des fonds communs de placement des clients.
2. Déterminez un intervalle de confiance à 95% de la valeur moyenne des fonds communs de placement des clients.

EXERCICE 5:

Dans la population française, le pourcentage d'individus dont le sang est de rhésus négatif est de 15%. Dans un échantillon représentatif de 200 Basques français on observe que 44 personnes sont de rhésus négatif.
Donner un intervalle de confiance à 99% de la proportion de Basques français ayant un rhésus négatif.

EXERCICE 6:

Les pignons d'une marque donnée pèsent en moyenne 0.5g avec un écart-type de 0.02g. Quelle est la probabilité pour que deux lots (choisis au hasard avec remise) de 1000 pignons chacun diffèrent entre eux, en moyenne, de plus de 0.2g ?

EXERCICE 7:

Les ampoules électriques d'un fabricant A ont une durée de vie moyenne μ_1 avec un écart-type $\sigma_1 = 200$h et celles d'un fabricant B ont une durée de vie moyenne μ_2 avec un écart-type $\sigma_2 = 100$h.
Un échantillon de 150 ampoules de A a donné une durée de vie moyenne de 1400h. Un échantillon de 100 ampoules B a donné une durée de vie moyenne de 1200h.
Déterminer un intervalle de confiance à 95% puis à 99 % de la différence des durées de vie moyenne des variétés A et B.

EXERCICE 8:

Dans une population, le pourcentage de fumeurs est de 60%. On tire au hasard un échantillon de 100 sujets. Quel risque y a t-il de perdre le pari que la proportion de fumeurs dans cet échantillon soit comprise entre 0.5 et 0.7 ?

EXERCICE 9:

Les parties A et B sont indépendantes.

A. Un atelier s'approvisionne avec des pièces produites en grande série. On note X, la variable aléatoire qui à toute pièce choisie au hasard dans la production associe sa masse en g. On admet que X suit une loi normale d'espérance 500g et d'écart-type σ.

1. On suppose dans cette question que σ est égal à 5g. Les pièces présentent le défaut A si leur masse n'est pas dans l'intervalle [495 ; 505]. On prélève au hasard une pièce dans la production. Quelle est la probabilité qu'elle présente le défaut A ?
2. Quelle doit-être la valeur de σ pour qu'une pièce de la production choisie au hasard présente le défaut A avec une probabilité inférieure à 0.05 ?

B. Les pièces de la production peuvent présenter un défaut B.

On veut estimer la proportion de pièces de la production présentant le défaut B par un intervalle de confiance. Pour cela, on prélève au hasard et avec remise un échantillon de 1000 pièces et on constate que 70 d'entre elles présentent le défaut B.

Déterminer un intervalle de confiance à 98% de la proportion de pièces de la production présentant le défaut B.

EXERCICE 10:

Une entreprise commercialise des pieds de lit de type boule. Pour ces pieds on utilise une bague en matière plastique de diamètre intérieur x. On définit ainsi une variable aléatoire X, qui à chaque bague tirée au hasard dans la production, associe son diamètre intérieur x mesuré en millimètres. On admet que X suit le loi normale de moyenne μ et d'écart-type 0,04. Le fournisseur affirme que $\mu = 12{,}1$.

On a un doute sur cette affirmation. On prélève un échantillon de 64 pièces dans la livraison. Le diamètre intérieur moyen sur cet échantillon est de 12,095.

Que concluez-vous au seuil de signification de 10% quant au diamètre intérieur moyen des bagues ?

EXERCICE 11 :

Une usine fabrique des câbles. Un câble est considéré comme conforme si sa résistance à la rupture est supérieure à 3 tonnes. L'ingénieur responsable de la production voudrait connaître, en moyenne, la résistance à la rupture des câbles fabriqués.

Il n'est, bien sûr, pas question de faire le test de rupture sur toute la production (l'usine perdrait toute sa production !). Notons X la variable aléatoire correspondant à la force à exercer sur le câble pour le rompre (en tonnes).

Un technicien prélève donc un échantillon de 100 câbles dans la production.

Avec les données de l'échantillon, le technicien obtient les résultats suivants : la résistance moyenne à la rupture des 100 câbles de l'échantillon est de 3.5 tonnes avec un écart-type de 0.4 tonne.

1. Décrire l'expérience aléatoire: ε
2. Décrire la population étudiée: Ω
3. Quelle probabilité utilisez-vous dans votre espace probabilisable ?
4. Décrire sur votre espace probabilisé la variable aléatoire étudiée: X
5. Que représente le paramètre μ ?
6. Donner un estimateur puis une estimation de μ.
7. Donner un estimateur puis une estimation de $\sigma^2 = \text{var}(X)$
8. Peut-on dire, avec un risque d'erreur de 2.5% que la résistance moyenne à la rupture de l'ensemble des câbles de la production est strictement supérieure à 3 tonnes ?
   Pour cette question, on supposera que la variable aléatoire $X \sim N(\mu, \sigma)$ et que la valeur de σ est ici connue et égale à 0.38.

La proportion de câbles dont la résistance est supérieure à 3 tonnes dans cet échantillon est de 0,85.

1. Décrire la nouvelle variable aléatoire étudiée : X'
2. Quelle est sa loi ?
3. Donner une estimation ponctuelle de la proportion π de câbles conformes dans la production.
4. Peut-on dire, avec un risque d'erreur de 5% que la proportion π de câbles conformes dans la production est strictement supérieure à 0.80 ?

EXERCICE 12 :

On utilise une nouvelle variété de pommes de terre dans une exploitation agricole.
Le rendement moyen de l'ancienne variété était de 41.5 tonnes à l'hectare.
La nouvelle variété est cultivée sur 100 hectares, avec un rendement moyen de 45 tonnes à l'hectare et un écart-type (échantillonnal) de 11.25.

1. Faut-il, avec un risque d'erreur de α=1%, favoriser la culture de la nouvelle variété ?
2. Calculer la puissance du test précédent si le « vrai » rendement moyen de la nouvelle variété est supposée égal à 44 tonnes. Qu'en pensez-vous ? Calculez alors le risque d'erreur de deuxième espèce.

EXERCICE 13 :

Un échantillon de 112 malades atteints d'un cancer du colon a été comparé à 185 témoins non malades quant à leur consommation moyenne de caféine. Pour les malades, cette consommation moyenne est égale à 147.2 mg/jour (l'écart-type échantillonnal est de 101.8 mg/j) et pour les témoins, elle vaut 132.9 mg/j (l'écart-type échantillonnal est de 115.7 mg/j). Tester, avec un risque de première espèce α = 5%, si la consommation moyenne de caféine diffère entre les malades et les non malades. Peut-on inférer une association entre la consommation de caféine et le cancer du colon ?

EXERCICE 14:

Pour un sondage électoral, on constitue deux échantillons d'électeurs de tailles 300 et 200 respectivement dans 2 circonscriptions A et B.
Cela met en évidence des intentions de vote de 56% et 48% pour un candidat donné.
Tester, au seuil de 5% les hypothèses suivantes:
1. Il y a une différence entre les circonscriptions
2. Le candidat est préféré dans la circonscription A

EXERCICE 15:

Une société de production d'électricité éolienne, cherche à comparer l'efficacité de deux types d'éoliennes : une éolienne à deux pales (E2p) et une éolienne à trois pales (E3p). Pour ce faire, elle a installé sur un même parc éolien une éolienne de chaque type, et a relevé les puissances de chaque éolienne (en kW) toutes les 10 minutes.
Afin de comparer les productions des éoliennes, l'ingénieur statisticien a prélevé aléatoirement dans la base de données, et ce de façon indépendante pour chaque éolienne, les 9 puissances (en kW) suivantes :

E2p	5	18	19	11	6	19	20	22	17
E3p	2	22	28	12	6	18	29	21	24

1. Définir clairement les deux variables aléatoires étudiées :
   X^1 (puissance de l'éolienne à 2 pales) et X^2 (puissance de l'éolienne à 3 pales)
2. Donner une estimation ponctuelle et un intervalle de confiance à 95% de la puissance moyenne de chaque éolienne.
   On notera μ_1 la puissance moyenne de l'éolienne à 2 pales et μ_2 la puissance moyenne de l'éolienne à 3 pales. On justifiera les hypothèses éventuellement nécessaires.
3. Donner une estimation ponctuelle de la variabilité de la puissance de chaque éolienne.
   On notera σ_1 l'écart-type de la variable X^1 et σ_2 l'écart-type de la variable X^2.
4. Peut-on supposer que les puissances des deux éoliennes ont la même variabilité ?
5. Peut-on affirmer, avec un risque d'erreur de 1%, que la puissance moyenne de l'éolienne à 3 pales est supérieure à la puissance moyenne de l'éolienne à 2 pales ?
6. Pouvez-vous, avec cette étude, conseiller à la société un type particulier d'éolienne ?

EXERCICE 16:

Montrer que la statistique S^{*2} est un estimateur sans biais de σ^2.
Rappel :

$$S^{*2}:\begin{cases} E_n \to \Re \\ e_n \mapsto s^{*2} = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2 \end{cases}$$

EXERCICE 17:

Soit un échantillon aléatoire $(X_1, \ldots, X_n)$ iid (indépendantes et identiquement distribuées) **extrait** d'une variable aléatoire $X \sim N(\mu, \sigma)$.

Démontrer, grâce au théorème de Fisher (généralisation), que:

$$RC = \frac{(\bar{X}_n - \mu)}{\frac{S^*}{\sqrt{n}}} \sim t_{(n-1)}$$

Avec:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$$

COMPLEMENTS SUR LES ESTIMATEURS
COURS

1/ INTRODUCTION

Nous avons vu que $\bar{X}_n$, S^2 et P_n sont respectivement des estimateurs de μ, σ^2 et de π (paramètres de la population).
Cependant un même paramètre peut être estimé à l'aide d'estimateurs différents : pour une distribution symétrique, la médiane de l'échantillon est également une estimation de μ.
Afin de choisir entre plusieurs estimateurs possibles d'un même paramètre il faut définir les qualités exigées d'un estimateur.

2/ QUALITES D'UN ESTIMATEUR

Soit θ le paramètre de la population à estimer et T un estimateur de θ.

2.1/ Estimateur convergent

La première qualité d'un estimateur est d'être convergent*.

On souhaite pouvoir, en augmentant la taille de l'échantillon, diminuer l'erreur commise en prenant la valeur observée de T à la place de θ. Si c'est le cas, on dit que l'estimateur est **convergent** (on voit aussi *consistant*), c'est-à-dire qu'il converge vers sa vraie valeur.

C'est le cas des estimateurs présentés en cours ($\bar{X}_n$, S^2 et P_n).
** : Noter qu'il existe différents types de convergence des suites de variables aléatoires (convergence en probabilité, convergence presque sûre ou convergence forte, convergence en moyenne d'ordre p et convergence en loi)*

Deux estimateurs convergents ne convergent cependant pas à la même vitesse, ceci est lié, pour une taille d'échantillon donnée, à la notion de **précision** d'un estimateur.
Un estimateur est une variable aléatoire. L'erreur d'estimation T - θ qui est une variable aléatoire se décompose de façon élémentaire en :
$T - E(T) + E(T) - \theta$ où E(T) est l'espérance de l'estimateur.
$T - E(T)$ représente les fluctuations aléatoires de T autour de sa valeur moyenne tandis que $E(T) - \theta$ est assimilable à une erreur systématique due au fait que T varie autour de sa valeur centrale E(T) et non autour de θ.
La quantité $E(T) - \theta$ s'appelle **le biais**.

2.2/ Estimateur sans biais

Il est donc souhaitable d'utiliser des estimateurs sans biais.

Définition :

T est un estimateur sans biais de θ si

$E(T) = \theta$

2.3/ Précision d'un estimateur

On mesure généralement la précision d'un estimateur T par l'erreur quadratique moyenne :

$E[(T-\theta)^2]$

On peut écrire :

$$E[(T-\theta)^2] = E[(T-E(T)+E(T)-\theta)^2]$$

$$= E[(T-E(T))^2] + 2E[(T-E(T))(E(T)-\theta)\ ] + E[(E(T)-\theta)^2]$$

Comme $E(T)-\theta$ est une constante et que $E[T-E(T)\ ] = 0$, il vient :

$$E[(T-\theta)^2] = V(T) - (E(T)-\theta)^2$$

Ainsi, de deux estimateurs sans biais, le plus précis est donc celui de variance minimale.

2.4/ Recherche du meilleur estimateur d'un paramètre θ

Ainsi, la précision d'un estimateur dépend de sa variance et celle-ci ne peut en général se calculer que si l'on connaît la loi de T qui dé pend de celle des X_i.

Le modèle utilisé en théorie classique de l'estimation est alors le suivant : on observe un échantillon d'une variable X dont on connaît la loi de probabilité à l'exception de la valeur numérique d'un ou de plusieurs paramètres (par exemple : X suit une loi de Poisson $P(\theta)$ de paramètre θ inconnu). En d'autres termes la variable X est définie par une famille paramétrée de lois $f(x,\theta)$ où f a une expression analytique connue.

De plus, la théorie de l'estimation ne permet pas de résoudre le problème de la recherche d'estimateurs d'erreur quadratique minimale. On se contentera de rechercher pour une famille de loi donnée $f(x,\theta)$ l'estimateur **sans biais** de θ **de variance minimale**.

Il reste cependant possible dans certain cas particulier de trouver des estimateurs biaisés plus précis que le meilleur estimateur sans biais.

3/ METHODES DE CONSTRUCTION D'UN ESTIMATEUR

3.1/ Méthode du maximum de vraisemblance (MV)

La méthode du maximum de vraisemblance consiste à choisir comme estimateurs des paramètres inconnus $\theta = (\theta_1, \theta_2, ..., \theta_k)$, les valeurs qui rendent maximum la probabilité d'avoir obtenu l'échantillon.

Soit X une variable aléatoire dont la fonction de densité $f(x, \theta)$ dépend du paramètre $\theta = (\theta_1, \theta_2, ..., \theta_k)$, et soit $(x_1, x_2, ..., x_n)$ une réalisation de l'échantillon aléatoire $(X_1, X_2, ..., X_n)$.

La fonction Vraisemblance est la fonction :

$$\theta \longrightarrow V(\theta) = f_{(X_1,...,X_n)}(x_1, ...x_n)$$
$$= f_{X_1}(x_1) * f_{X_2}(x_2) * ..... * f_{X_n}(x_n)$$

= fonction de densité du n-uplet de variables aléatoires $(X_1, X_2, ..., X_n)$ en $(x_1, x_2, ..., x_n)$ car les (X_i) sont indépendants (donc produit des fonctions de densité des (X_i))

Dans le cas d'une loi discrète, c'est la fonction :

$$\theta \longrightarrow V(\theta) = P((X_1 = x_1) \cap (X_2 = x_2) \cap ... \cap (X_n = x_n))$$
$$= P(X_1 = x_1) * P(X_2 = x_2) * ..... * P(X_n = x_n)$$

Un estimateur du maximum de vraisemblance (EMV) $\hat{\theta}$ vérifie donc :

$$\forall \theta, \quad V(\theta) \leq V\left(\hat{\theta}\right)$$

$V(\theta)$: la fonction Vraisemblance de θ

Lorsque f > 0, il est équivalent et généralement plus facile de chercher le maximum de la fonction log-vraisemblance de θ : $\ln[V(\theta)]$ (attention : il s'agit du logarithme népérien : ln).

On note aussi $\ln[V(\theta)] = LV(\theta)$

$LV(\theta)$: la fonction Log-vraisemblance de θ

Théorème :

Si f > 0 et si $\frac{\partial f}{\partial \theta}$ et $\frac{\partial^2 f}{\partial \theta^2}$ sont continues, l'estimateur $\hat{\theta}$ du maximum de vraisemblance vérifie :

1. Les équations de vraisemblance :

$$\forall i \in \{1, 2, ...k\}; \quad \frac{\partial LV}{\partial \theta_i}\left(\hat{\theta}\right) = 0$$

2. La condition suffisante de maximum :
   La matrice hessienne de LV ,
   $\left(\frac{\partial^2 LV}{\partial\theta_i\partial\theta_j}(\hat{\theta})\right)_{1\le i,j\le k}$ est définie négative

   Remarque:

   *Soit M une matrice symétrique réelle d'ordre n. Elle est dite **définie positive** si elle vérifie la propriété suivante :*

   *Pour tout vecteur colonne **non nul** $\mathbf{x}$ à n éléments réels, on a :*

   $x'Mx > 0$

   Noter que : $x'Mx = \langle x, x\rangle_M = \|x\|_M^2$

   *Elle est dite **définie négative** si son opposée (-M) est définie positive !*

A retenir:

$V(\theta)$ est donc soit la densité de $(X_1, X_2, ..., X_n)$ si X est absolument continue, soit la probabilité conjointe $P(X_1 = x_1 \cap X_2 = x_2 ..... \cap X_n = x_n)$ si X est discrète.

$V(\theta)$ considéré comme fonction de θ seul est appelé la « vraisemblance ».

On appelle EMV : Estimateur du Maximum de Vraisemblance

3.2/ Méthode des moments

Cette procédure d'estimation est, à première vue, totalement empirique, mais pleine de bon sens. En fait, elle repose sur **la propriété de convergence presque sûre des moments empiriques d'un échantillon i.i.d.** $(X_1, X_2, ..., X_n)$ **extrait de X, vers les moments théoriques correspondants de X.**

Soit $\theta = (\theta_1, \theta_2, ..., \theta_k)$, nous noterons par $m_p(\theta)$ le moment théorique d'ordre p de X, qu'il soit centré ou non, et par $m_p(e_n)$ le moment empirique d'ordre p de X.

Définition:

On appelle estimateur de θ obtenu par la méthode des moments (EMM) la solution du système :

$m_1(\theta) = m_1(e_n)$

.

$m_k(\theta) = m_k(e_n)$

Remarque :
Le choix des moments est guidé par la facilité de résolution du système. On peut prendre des moments tous centrés en E(X), ou tous non centrés, ou un mélange de moments centrés ou non centrés. En outre, il n'y a aucune raison de retenir les k premiers moments, sinon la simplicité de calcul.
On appelle EMM : Estimateur de la Méthode des Moments

Rappel:
Soit X une variable aléatoire, on appelle, s'ils existent ::

- le moment **théorique** d'ordre p de X :
  $E(X^p)$
- le moment **centré théorique** d'ordre p de X :
  $E\left[\left(X - E(X)\right)^p\right]$

Si l'on observe n réalisations $\left(x_1, x_2, ..., x_n\right)$ de X , on appelle :

- le moment **empirique** d'ordre p de X
  $\frac{1}{n}\sum_{i=1}^{n}\left(x_i\right)^p$
- le moment **centré empirique** d'ordre p de X

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}_n\right)^p \quad \text{avec} \quad \overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

EXERCICES

EXERCICE 18:

Considérons un échantillon aléatoire $(X_1, X_2, ..., X_n)$ issu d'une variable aléatoire parente

$X \sim N(\mu, \sigma)$

1. Donner les estimateurs du maximum de vraisemblance (EMV) de μ et de σ^2
2. Donner les estimateurs par la méthode des moments (EMM) de μ et de σ^2
3. Que constatez-vous ?
4. L'estimateur du maximum de vraisemblance (EMV) de μ est –il sans biais ?
5. L'estimateur du maximum de vraisemblance (EMV) de σ^2 est –il sans biais ?

EXERCICE 19:

Les oiseaux d'un certain type prennent leur envol après avoir effectué quelques sauts sur le sol. On suppose que ce nombre X de sauts peut être modélisé par une distribution géométrique :

$P(X = x) = p(1-p)^x \quad x \geq 0$

Pour n=130 oiseaux de ce type, on a relevé les données suivantes :

Nombre de sauts x	1	2	3	4	5	6	7	8	9	10	11	12
Fréquence de x	48	31	20	9	6	5	4	2	1	1	2	1

1. Quel est l'estimateur du maximum de vraisemblance (EMV) de p ?
2. Calculer la valeur de cet estimateur avec les données de l'échantillon

EXERCICE 20:

Soit $X \sim \gamma(p, \theta)$ loi gamma

Trouver les estimateurs de p et de θ par la méthode des moments

EXERCICE 21:

Statistique bayésienne

EXERCICE 22:

Les données ci-dessous ont été observées sur un échantillon aléatoire de 500 travailleurs. Pour chacun, on a évalué le niveau de stress au travail et mesuré le temps pris pour se rendre au travail.
Pour établir le tableau de contingence, on a regroupé les valeurs de la variable « temps » » pour créer 3 catégories. Voici les résultats obtenus :

Temps pour se rendre au travail *Niveau de stress*	*Moins de 15minutes*	*De 15 à 45 minutes*	*Plus de 45 minutes*	*Total*
Elevé	32	73	58	163
Modéré	29	34	36	99
Fort	77	121	40	238
Total	138	228	134	500

Ces données permettent-elles de conclure qu'il existe une relation significative entre le niveau de stress des travailleurs et le temps qu'ils prennent pour se rendre au travail?
On prendra un risque d'erreur de première espèce de 5 %