Université Mohammed Sedik Benyahia-Jijel

Faculté des Sciences Exactes et Informatique

Département de Mathématiques

STATISTIQUE INFÉRENTIELLE

RÉSUMÉ DU COURS

Niveau : Troisième Anneé

Spécialité : Mathématiques

Année universitaire 2020/2021

Enseignant : GHERDA Mebrouk

Table des matières

Chapitre 1

MODES DE CONVERGENCES ET APPROXIMATIONS

1.1 MODES DE CONVERGENCES

Soit X une varible alétoire et (X_n) une suite de variables alétoires définies sue le même espace probabilisé $(\Omega,\ \Lambda,\ P)$.

1.1.1 Convergence en probabilité : $X_n \longrightarrow_P X$

Définition 1.1. *la suite* (X_n) *Converge en probabilité vers* X *si* $\forall \varepsilon > 0\ \lim_{n\longrightarrow\infty} P\left(|X_n - X| > \varepsilon\right) = 0$

La loi faible des grand nombres

Si les variables aléatoires X_n sont deux à deux non covariées, de même loi, d'espérance μ de variance σ^2, alors $\frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow_P \mu$.

1.1.2 Convergence en moyenne quadratique : $X_n \longrightarrow_{mq} X$

Définition 1.2. *la suite* (X_n) *Converge en moyenne quadratique vers* X *si* $\lim_{n\longrightarrow\infty} P\left(\left(X_n - X\right)^2\right) = 0$

2.2 Propriétés : (X_n) Converge en moyenne quadratique vers X si et seulement si $\lim_{n\longrightarrow\infty} E\left(X_n\right) = E\left(X\right)$ et $\lim_{n\longrightarrow\infty} var\left(X_n - X\right) = 0$.

1.1.3 Convergence presque sûre : $X_n \longrightarrow_{ps} X$.

Définition 1.3. *la suite* (X_n) *Converge presque sû rement vers* X *si* $P\left(w/\lim_{0\to\infty} X_n\left(w\right) = X\left(w\right)\right) = 1$

Loi forte des grands nombres

Si les variables aléatoires X_n sont mutuellement indépendantes de même loi, d'espérance μ, alors $\frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow_{Ps} \mu$.

1.1.4 Convergence en loi : $X_n \longrightarrow_l X$

Soit F_n la fonction de répartition de X_n et F celle de X.

Définition 1.4. *la suite* (X_n) *Converge en loi vers* X *si por tout* x *où* F *est continue,* $\lim_{n\to\infty} F_n(x) = F(x)$.

Distance de Kolmogorov entre deux fonction de répartition G_1 et G_2. Elle est définit par $\Delta(G_1, G_2) = \sup_{x\in\mathbb{R}} |G_1(x) - G_2(x)|$.

Proprités de la convergence en loi

-Si $\lim_{n\to\infty} \Delta(F_n, F) = 0$ alors $X_n \longrightarrow_l X$.

-Si F est continue alors : $X_n \longrightarrow_l X$ si et seulement si $\lim_{n\to\infty} \Delta(F_n, F) = 0$.

-Si X_n et X sot des variables aléatoires à valeurs dans $\mathbb{N}$ alors $X_n \longrightarrow_l X$ si et seulement si : $\forall k \in \mathbb{N}$, $\lim_{n\to\infty} P(X_n = k) = P(X = k)$.

-Soit a et b deux réels. Si $X_n \longrightarrow_l X$ alors $aX_n + b \longrightarrow_l aX + b$.

1.1.5 Théorème limite centrale

Si les variables aléatoires X_n sont mutuellement indépendantes de même loi, d'espérance μ et d'écart-type σ diff érent de 0 alors :

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right) \longrightarrow_l X$$

où X est une variable aléatoire de loi de Laplace-Gauss centré e réduite.

1.1.6 Liens entre les différents type de convergence.

Ils se résume de la façon suivante :

$X_n \longrightarrow_{Ps} X \Longrightarrow X_n \longrightarrow_P X \Longrightarrow X_n \longrightarrow_l X$

$X_n \longrightarrow_{mq} X \Longrightarrow X_n \longrightarrow_P X \Longrightarrow X_n \longrightarrow_l X$

La convergence en loi est la seule qui ne fait intervenir que les lois des variables aléatoires.

Dans le cas où X est une variables aléatoire égale à a ou presque sûrement égale à a :

$$X_n \longrightarrow_P X \Leftrightarrow X_n \longrightarrow_l X$$

1.2 APPROXIMATION

1.2.1 Approximation d'une loi binomiale par une loi de Poisson

Considéronons une loi binomiale de paramètres n et p ; Si n et grand et p assez petit, la loi de Poisson est une bonne approximation de la loi binomiale à condition que le produit np reste fini, et dans ce cas la loi binomiale $B(n,p)$ tend vers la loi de Poisson $P(\lambda = np)$

En pratique, nous utiliserons l'approximation de la loi binomiale par la loi de Poisson dans les conditions suivantes :

a) $n > 50$, $p < 0.1$ $n > 50$, $p > 0.9$ car alors $q < 0.1$ ce qui nous ramène au cas précédent compte tenu du rôle symétrique que jouent p et q dans le cas d'une loi binomiale.

1.2.2 Approximation d'une loi binomiale par une loi normale

Soit une variable aléatoire discrète X suivant une loi binomiale $B(n,p)$ telle que : $P(X = k) = C_n^k p^k q^{n-k}$.

Si n est suffisamment grand et p pas trop proche de 0 ni de 1 avec $np \geqslant 5$ et $nq \geqslant 5$ alors la loi normale de paramètres $m = np$ et $\sigma = \sqrt{npq}$ constitue une bonne approximation de la loi binomiale.

Remarque 1. *Il y a nécessité de remplacer $P(X = k)$ par $P(k - 0.5 < X < k + 0.5)$ (correction de continuité) car dans le cas d'une loi discrète les probabilités de type $P(X = k)$ sont nulles.*

Les conditions pratique de l'approximation sont :

$n \geqslant 30$ $p \in [0.1, 0.9]$ car sinon la loi de Poisson réalise une meilleure approximation

$np \geqslant 5, nq \geqslant 5$.

1.2.3 Approximation de loi de Poisson par une loi normale

Soit une variable aléatoire discrète X suivant la loi de Poisson $P(\lambda)$ telle que : $P(X = k) = \dfrac{\lambda^k}{k!} \varrho^{-\lambda}$

Si λ est suffisamment grand, la loi normale de paramètres $m = \lambda$ et $\sigma = \sqrt{\lambda}$ constitue une bonne approximation de la loi de Poisson..la correction de continuité citée ci-dessus s'applique ici aussi et $P(X = k) = P(k - 0.5 < X < k + 0.5)$

APPLICATION

On sait que la probabilité qu'une personne soit allergique à un certain médicament est égale à $(10)^{-3}$, On s'int éresse à un échantillon de 1000 personnes. On appelle X la variable aléatoire dont la valeur est le nombre de personne allergique dans l'échantillon.

1-Déterminer, on la justifiant, la loiu de probabilité de X.

2-En utilisant une approximation que l'on justifiera, calculer les probabilit és des événements suivants :

a-Il y a exactement deux personnes allergique dans l'echantillon

b- Il y a au moins deux personnes allergiques dans l'échantillon.

Que peut-on dire si 30% de la population d'où provient cet é chantillon sont allergiques à ce médicament.

1.2.4 Lois dérivées de la loi normale

Loi du khi carré : χ_n^2

$\chi_n^2 : z_1^2 + z_2^2 + ... + z_n^2 = \chi_n^2$

En particulier, $\chi_1^2 = [N(0,1)]^2$.

Limite de la loi du khi-carré

Quand n devient grand [en pratique, quand $n \geq 30$], la loi du χ^2 tend vers une nouvelle loi normale $N(m, s)$ de moyenne $m = n$ et d'é cart type $\sigma = \sqrt{2n}$. Il suffit de centrer et réduire pour passer de la loi du χ^2 à une loi normale centrée ré duite z.

Par conséquent, $\dfrac{\chi_{n,\alpha}^2 - n}{\sqrt{2n}} = z_\alpha$ ou inversement, $\chi_{n,\alpha}^2 = n + z_\alpha\sqrt{2n}$

pour toute valeur de probabilité α.

Remarque 2. *L'analyse des données biologiques utilise abondamment la loi du* χ^2

Loi de Fisher-Snedecor :

Définition 1.5. $F_{(v_1,v_2)} = \dfrac{\chi_{v_1}^2/v_1}{\chi_{v_2}^2/v_2}$

Le rapport de deux variables aléatoires distribuées comme khi-carr é, chacune divisée par ses degrés de liberté, est une variable aléatoire distribuée comme F.

Il existe autant de courbes de densité de probabilité de F que de

combinaisons possibles de n_1 et n_2.

Applications

• Test F de rapport de variances.

• Analyse de variance.

1.2.5 Loi de Student : t(n)

Loi décrite en 1908 par William Sealy Gosset sous le pseudonyme "Student". Le premier article de Student, publié en 1907, avait établi que la distribution des dénombrements de cellules dans les carrés d'un hémacytomètre suivaient la loi de Poisson (répartition aléatoire).

Deux définitions équivalentes de la loi de t :

1) $t(n) = \dfrac{Z}{\sqrt{\chi^2/n}}$

2) $t(n) = F_{(v_1,v_2)}$ lorsque $n_1 = 1$. Le nombre de degr és de liberté de la loi de t est alors $n = n_2$.

Applications

• Estimation des paramètres d'une population à partir de renseignements portant sur un échantillon.

• Test de comparaison des moyennes.

• Calcul de la probabilité d'observer un écart donné à la moyenne, en particulier dans le cas de petits échantillons :

Pour un écart observé , la probabilité d'une telle observation x_i est donnée par la variable aléatoire $t = (x_i - \overline{x})/s_x$.

Il existe autant de courbes de densité de probabilité de t que de valeurs possibles de n. Voir la table de la distribution de t

Théorème de la limite centrée

Soit (X_n) une suite de variables aléatoires mutuellement indé pendantes de même loi de moyenne μ et d'écart-type σ et soit $\overline{X} = \dfrac{1}{n}\left(\sum_{i=1}^{n} X_i\right)$. Pour $n \geq 30$, la variable aléatoire $\overline{X}$ suit, approximativement, la loi normale de moyenne μ et d'écart-type $\frac{\sigma}{\sqrt{n}}$.

Chapitre 2

Le modèle statistique

2.1 Notions et définitions

2.1.1 Le modèle statistique

Un modèle statistique est un objet mathématique associé à l'observation de données issues d'un phénomène aléatoire.

Une expérience statistique consiste a recueillir une observation x d'un elément aléatoire X, à valeurs dans un espace χ et dont on ne connait pas exactement la loi de probabilité P. Des consid érations de modélisation du phénomène observé amè nent à admettre que P appartient a une famille P de lois de probabilit é possibles.

Définition 2.1. *Le modèle statistique (ou la structure statistique) associé a cette expérience est le triplet* $(\chi; A; P)$*, où :*

X est l'espace des observations, ensemble de toutes les observations possibles.

A est la tribu des evènements observables associée.

P est une famille de lois de probabilit es possibles défnie sur A.

L'int erêt de cette notion de modèle statistique est qu'elle permet de traiter avec le même formalisme tous les types d'observations possibles.

On dit que le modèle est discret quand X est fini ou dénombrable. Dans ce cas, la tribu A est l'ensemble des parties de X : $A = P(X)$.

On dit que le modèle est continu quand X$\subset$ IR^p et $\forall$ P $\in P$, P admet une densité (par rapport à la mesure de Lebesgue) dans IR^p. Dans ce cas, A est la tribu des boréliens de X (tribu engendrée par les ouverts de X) : $A = B(X)$.

On peut aussi envisager des modèles ni continus ni discrets, par exemple si l'observation

a certains eléments continus et d'autres discrets. X et A sont alors plus complexes.

Le cas le plus fréquent, est celui o u l'élément aléatoire observé est constitué de variables aléatoires indépendantes et de même loi (*i.i.d.*) : $X = (X_1, ..., X_n)$, où les X_i sont *i.i.d.* On dit que l'on a alors un modèle d'échantillon.

Dans ce cas, par convention, si on note $(X; A; P)$ le modèle correspondant a un echantillon de taille 1, on notera $(X; A; P)^n$ le mod éle correspondant a un echantillon de taille n.

Exemple 2.2. *l'expérience consiste à recueillir les durées de vie, supposé es indépendantes et de même loi exponentielle, de n ampoules electriques. L'observation est de la forme $x = (x_1, ..., x_n)$, où les x_i sont des réalisations de variables al eatoires X_i indé pendantes et de même loi exponentielle de paramètre inconnu.*

Pour tout i, $x_i \in IR+$, donc l'espace des observations est $X = IR_+^n$. Alors la tribu associée est $A = B(IR_+^n)$. Le modè le est continu. Comme on admet que la loi est exponentielle mais que son paramètre est inconnu, l'ensemble des lois de probabilités possibles pour chaque X_i est $exp(\lambda)$; $\lambda \in IR_+$, Comme les X_i sont ind ependantes, la loi de probabilité du vecteur $(X_1, ..., X_n)$ est la loi produit $P = \{exp(\lambda)^{xn}; \lambda \in IR_+\}$, ensemble des lois de probabilité des vecteurs al éatoires de taille n dont les composantes sont indépendantes et de même loi exponentielle de paramètre inconnu.

Finalement, le modèle statistique associé est :

$(IR_+^n; B(IR_+^n); exp(\lambda)^{xn}; \lambda \in IR_+)$ qu'on peut aussi écrire : $(IR_+^n; B(IR_+^n); exp(\lambda); \lambda \in IR_+)^n$

Modèle paramétrique ou non paramétrique

Un modèle paramétrique est un modèle où l'on suppose que le type de loi de X est connu, mais qu'il dépend d'un paramètre inconnu, de dimension d. Alors, la famille de lois de probabilité possibles pour X peut s'écrire $P = \{p_\theta; \theta \in IR^d\}$.

Un modèle non paramétrique est un modèle où P ne peut pas se mettre sous la forme ci-dessus. Par exemple, P peut être :

l'ensemble des lois de probabilité continues sur IR,

l'ensemble des lois de probabilit e sur IR symétriques par rapport a l'origine, etc...

Dans ce cadre, il est possible de déterminer des estimations, des intervalles de confiance, d'effectuer des tests d'hypothèses. Mais les objets sur lesquels portent ces procédures statistiques ne sont plus des paramètres de lois de probabilité. On peut vouloir estimer des quantités réelles comme l'espérance et la variance des observations. On peut aussi vouloir estimer des fonctions, comme la fonction de répartition et la densité des observations.

2.1.2 Fonction de vraisemblance

Dans un modèle paramétrique, la fonction de vraisemblance joue un rôle fondamental.

Pour un modèle d'échantillon discret, l'élément aléatoire observé est $X = (X_1, ..., X_n)$, où les X_i sont ind ependantes et de même loi discrète. Alors la fonction de vraisemblance est :

$L(\theta; x_1, ..., x_n) = P(X_1 = x_1, ..., X_n = x_n; \theta) = \Pi_{i=1}^{n} P(X_i = x_i; \theta)$

Pour un modèle d'échantillon continu, l'élément aléatoire observé est $X = (X_1, ..., X_n)$, où les X_i sont ind ependantes et de même loi continue. Alors la fonction de vraisemblance est :

$L(\theta; x_1, ..., x_n) = f_{(X_1,...,X_n)}(x_1, ..., x_n; \theta) = \Pi_{i=1}^{n} f_{X_i}(xi; \theta)$

Définition 2.3. *La fonction de vraisemblance du modèle* $(X; A; \{P_\theta\}; \theta \in \Theta)$ *est la fonction de définie par :*

$\forall A \in A;\ P_\theta(A) = P(X \in A; \theta) = \int_A L(\theta; x) du(x)$ *:*

Plus généralement, pour toute fonction φ *intégrable, on a :* $E[\varphi(X)] =_\chi \varphi(x) L(\theta; x) du(x)$

Cas des modèles continus. Si X *est un vecteur alèatoire admettant une densitè* $f_X(x; \theta)$ *(par rapport a la mesure de Lebesgue), on sait bien que* $P(X \in A; \theta) = \int_A f_X(x; \theta) dx$.

Donc la mesure dominante est la mesure de Lebesgue et la fonction de vraisemblance est $L(\theta; x) = f_X(x; \theta)$.

Cas des modèles discrets. Si X est un vecteur aléatoire de loi discrète, définie par les probabilités elémentaires $P(X = x; \theta)$, alors :$P(X \in A; \theta) = \sum_{x \in A} P(X = x; \theta) = \int_A P(X = x; \theta) du_d(x)$

où u_d est la mesure de dénombrement sur X : $u_d(A) = card(A)$ et $\int_A f(x) du_d(x) = \sum_{x \in A} f(x)$. Donc la fonction de vraisemblance est bien $L(\theta; x) = P(X = x; \theta)$.

2.1.3 Statistique

Définition 2.4. *Dans un modèle statistique* $(X; A; P)$*, une statistique est une application mesurable* t *de* $(X; A)$ *dans un espace* Y *muni d'une tribu* B.

Définition 2.5. *La loi de probabilité* P_T *de* T *est appelée loi image par* t *et le modèle* $(Y; B; \{P_T; P \in P\})$ *est le modèle image par* t *de* $(X; A; P)$.

Exemple des ampoules. Le modèle est $(IR+; B(IR+); \{exp(\lambda); \lambda \in IR+\}^n$. $X = (X_1, ..., X_n)$,

où les X_i sont des variables aléatoires indépendantes et de même loi $exp(\lambda)$. On sait qu'alors $T = \sum_{i=1}^{n} X_i$ est de loi gamma $G(n; \lambda)$. Donc la loi image par $t(x) = \sum_{i=1}^{n} x_i$ est la loi $G(n; \lambda)$ et le modèle image est le mod ele $(IR+; B(IR+); \{G(n; \lambda); \lambda \in IR+\}$.

2.1.4 Modèle d'échantillonnage

Définition 2.6. *Soit une propriété définie par la v.a. X à valeur dans $\mathbf{X}$, application mesurable de $(\Omega, A, P) \to (X, B, P^X)$, B étant ici la tribu des Boréliens. Le modèle d'échantillonnage de taille n est l'espace produit $(\mathbf{X}, B, P)^n = (\mathbf{X}^n, B_n, P_n^X)$*

où - $\mathbf{X}^n = \mathbf{X} \times .... \times \mathbf{X}$ n fois est le produit cartésien de l'espace $\mathbf{X}$,

– B_n est la tribu produit des événements de $\mathbf{X}^n$,

– P_n^X est la loi ou la distribution jointe des observations.

On notera X_i la ième observation, v.a. de même loi que X et l'ensemble des observations $(X_1, ..., X_n)$ est l'échantillon aléatoire.

On notera que

$X_1, ..., X_n$ iid de loi P^X ou $(iid) \rightsquigarrow F_X$, F_X étant la fonction de répartition de X.

dans le cas où

la loi P^X est une loi discrète :

$$P(X_1 = x_1, ..., X_n = x_n) = \Pi_{j=1}^n P(X_j = x_j) = \Pi_{j=1}^n p_X(x_j)$$

ou la densité jointe dans le cas continu (P^X admet une densité f_X relativement à la mesure de Lebesgue) :

$$f_{(X_1,...,X_n)}(x_1, ..., x_n) = \Pi_{j=1}^n f_{X_j}(x_j) = \Pi_{j=1}^n f_X(x_j)$$

Cas de la population finie

On se place dans le cas d'une population E de taille finie N pour laquelle la propriété X n'est observée que sur un ensemble E_n de taille $n \leq N$. On note $(x_1, ..., x_N)$ l'ensemble des valeurs prises par la propriété X sur l'ensemble de la population $E = \{e_1, ..., e_N\}$. Ces valeurs sont déterministes, elles appartiennent à $\mathbf{X}$. On a alors les vraies moyenne μ et variance σ^2 de X :

$$\mu = \frac{1}{N} \sum_{j=1}^{N} .x_j; \qquad \sigma^2 = \frac{1}{N} \sum_{j=1}^{N} . (x_j - \mu)^2$$

La moyenne empirique

La moyenne empirique de l'échantillon est donnée par l'expression

$$\overline{X_N} = \mu = \frac{1}{N} \sum_{j=1}^{N} .X_j$$

Pour calculer $E(\overline{X_N})$ et $Var(\overline{X_N})$ dans le cas d'une population finie E de taille N, il faut distinguer le mode de tirage.

a. Tirage avec remise On a

$$E(\overline{X_n}) = \frac{1}{n}\sum_{j=1}^{n}.E(X_j)$$

Chacune des variables X_j est tirée de l'ensemble $\{x_1, ..., x_N\}$ avec la probabilité $1/N$, c'est-à-dire $P(X_j = x_l) = 1/N$, $\forall l = 1, ..., N$. D'où

$$E(X_j) = \frac{1}{n}\sum_{j=1}^{n}.x_l = \mu$$

(la vraie moyenne de la population) et $E(\overline{X_N}) = \mu$.

Pour calculer la variance, notons que les Xj sont des variables aléatoires indépendantes, et donc

$Var(\overline{X_n})$=$\frac{1}{n^2}\sum_{j=1}^{n}.Var(X_j)$

2.1.5 Familles Exponentielles

Définition 2.7. *Un modèle paramétrique important en Statistique est celui des familles exponentielles. Il recouvre de nombreux modèle paramétriques classiques : normal, binomial, poisson, gamma etc...*

. Un modèle statistique $(E; E; P)$ sur un espace des observations E est dit famille exponentielle générale s'il existe un entier p, des fonctions η,T ,C et h tels que les densités puisse s'écrire, pour tout θ de Θ, sous la forme : $f_\theta(x) = e^{<\eta(\theta),T(x)>}C(\theta)h(x)$;

avec les contraintes que T soit une fonction mesurable à valeurs dans R^p ; η soit une fonction à valeurs dans R^p ; C soit une fonction réelle positive qui ne dépend pas x ; h soit une fonction borélienne positive qui ne dépend pas de θ. Le vecteur aléatoire $T(X)$ est appelé statistique canonique du modèle. Si la fonction T est l'identité, la famille exponentielle est dite naturelle. On parle de forme canonique d'une famille exponentielle générale quand les densités de probabilités ont la forme $f_\theta(x) = e^{<\theta,T(x)>}C(\theta)h(x)$; pour tout θ de Θ, ce qu'il est toujours possible d'obtenir quitte à reparamétriser la famille par $\theta' = \eta^\theta$. Dans ce cas le paramètre θ de la famille exponentielle est appelé paramètre canonique.

Exemple 2.8. *Revenons sur le modèle de Bernoulli. La densité s'écrit :* $f_p(x) = p^x(1-p)^{1-x} = (\frac{p}{1-p}^x)(1-p)$
$= e^{xln(\frac{p}{1-p})}(1-p) = e^{<\eta(p),T(x)>}C(p)h(x)$ *;*

avec $\eta(p) = \frac{p}{1-p}; T(x) = x; C(p) = (1p)eth(x) = 1.$

2.1.6 Modèle position-échelle

Considérons un vecteur aléatoire X de loi P connue sur $(\mathbb{R}^n; B_{\mathbb{R}^n})$net A un sous espace de $\mathbb{R}^n$. Pour tout a dans A et tout b dans $\mathbb{R}_+$, on note $P_{a;b}$ la loin du vecteur $Y = a + bX$.

$P_{A;b} = \{P_{a;b} : a \in A; b \in \mathbb{R}_+\}$ est appelé modèle position-échelle engendré par P (ou par X). Le paramètre a est appelé paramètre de position et b paramètre d'échelle.

Si b est fixé (par exemple à 1) on parle de modèle de position. Dans le cas où A ne contient que le vecteur nul de $\mathbb{R}^n$, on parle de modèle échelle. **Exemple :Le Modèle gaussien unidimensionne**

Le modèle $P = \left\{N(u; \sigma^2); u \in \mathbb{R}\right\}$ est un modèle position engendré par la loi $N(0; \sigma^2)$.

Le modèle $P = \left\{N(u; \sigma^2); u \in \mathbb{R}, \sigma^2 > 0\right\}$ est un modèle position-échelle engendré par la loi $N(0; 1)$.

2.2 Exhaustivité

Statistique

Soit X une v.a. à valeurs dans $(\mathbf{X}, B)$ et soit $(\mathbf{Y}, C)$ un espace mesurable auxiliaire quelconque.

Définition 2.9. *On appelle statistique toute application T mesurable de $\mathbf{X}^n$ dans $\mathbf{Y}$, $\forall\ n\ T : \mathbf{X}^n \to \mathbf{Y}$*

Par exemple, $\mathbf{X} = \mathbf{Y} = \mathbb{R}$ et

$$T(X_1, ..., X_n) = \frac{1}{n}\sum_{j=1}^{n} X_j = \overline{X_n}$$

$$T(X_1, ..., X_n) = \frac{1}{n}\sum_{j=1}^{n} \left(X_j - \overline{X_n}\right)^2$$

ou $\mathbf{X} = \mathbb{R}$, $\mathbf{Y}= \mathbb{R}^n$ et $T(X_1, ..., X_n) = (X_{(1)}, ..., X_{(n)})$, où $X_{(1)} \leq X_{(2)}...X_{(n)}$

(cette statistique porte le nom de statistique d'ordre.

2.2.1 Statistique exhaustive

Définition 2.10. *On appelle modèle statistique paramétrique de paramètre $\theta \in \Theta$ pour un certain espace de dimension fini le couple $(\mathbf{X}, P_\theta)$, où $\mathbf{X}$ est l'espace des valeurs de X, v.a. du modèle, et P_θ la loi de probabilité de X*

Définition 2.11. *La statistique T sera dite **exhaustive pour** θ si la loi conditionnelle de X sachant $T(X) = t$ n'est pas une fonction du paramètre θ : $P_\theta(X|T(X) = t)$ ne dépend pas de θ*

On notera $f(x, \theta)$ la densité de P_θ relativement à une mesure dominante et
$sigma-$finie, μ. On va se restreindre au cas où μ est la mesure de Lebesgue (variables aléatoires de loi absolument

continue) et on retrouve la densité $f_\theta(x)$ ou la mesure de comptage (variables aléatoires de loi discrète) et on retrouve le système $P_\theta(X = x)$. On note X l'échantillon $(X_1, ..., X_n)$ issu du même modèle (X, P_θ)

Exemple la vraisemblance d'un échantillon $X = (X_1; ...; X_n)$ dans un tel modèle est :

$$L(x_1, ..., x_n; p) = P^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

On peut écrire :

$L(x_1, ..., x_n; p) = g_p(T(\underline{x}))h(\underline{x})$;

avec $g_p(x) = p^x(1-p)^{n-x}$ et h égale à 1. Grâce au théorème de factorisation on retrouve que la Statistique $T(X) = \sum_{i=1}^{n} X_i$ est bien exhaustive pour le paramètre p dans ce modèle.

Théorème : (Théorème de factorisation) Soit le modèle $(X, P_\theta$ et T une statistique $(\mathbf{X}^n, B_n) \to (\mathbf{Y}, C)$. T est exhaustive pour θ si et seulement s'il existe deux fonctions mesurables $g : \mathbf{X} \to \mathbb{R}_+$ et $h : \mathbf{Y} \to \mathbb{R}_+$ telles que $f(x, \theta)$ se met sous la forme $f(x, \theta) = h(x)g(T(x), \theta)$ où $x = (x_1, ..., x_n)$.

Exemples :

– Soit $X \sim U[0, \theta]$. On a $f(x_1, ..., x_n, \theta) = \frac{1}{\theta^n}\mathbf{1}_{sup1\leq j\leq nxj\leq\theta}$

En posant $h(x) = 1$ et $g(T(x), \theta) = \frac{1}{\theta^n}\mathbf{1}_{T(x)\leq\theta}$

on déduit que $T : x \longmapsto sup_{1\leq j\leq n}x_j$ est une statistique exhaustive pour θ.

Soit $X \sim exp(\theta)$. On a $f(x_1, ..., x_n, \theta) = \frac{1}{\theta^n}\exp\left(-\theta\sum_{j=1}^{n} x_j\right)$

et donc $T(X_1, ..., X_n) = \sum_{j=1}^{n} X_j$ est bien une statistique exhaustive. Soit $X \sim P(\theta)$. On a $f(x_1, ..., x_n, \theta) = e^{-n\theta}\theta^{\sum_{j=1}^{n} x_j}\Pi_{j=1}^{n}x_j!$

et donc $T(X_1, ..., X_n) = \sum_{j=1}^{n} X_j$ est bien une statistique exhaustive.

– Soit $X \sim N(\mu, \sigma^2)$. Alors la statistique $T(X_1, ..., X_n) = \left(\frac{1}{n}\sum_{j=1}^{n} X_j; \frac{1}{n}\sum_{j=1}^{n} X_j^2\right)$ est une statistique exhaustive pour $\theta = (\mu, \sigma^2)$.

2.2.2 Notion d'identifiabilité

Soit $(\mathbf{X}, P_{\theta)})$, $\theta) \in \Theta$ un modèle statistique paramétrique.

Définition 2.12. *Une valeur du paramètre $\theta_0 \in \Theta$ est identifiable si $\forall\theta) \neq \theta)_0$, $P_{\theta)} \overline{\neq P_{\theta_0}}$.* Le modèle $(\mathbf{X}, P_{\theta)})$, $\theta \in \Theta$ est dit identifiable si tous les paramètres sont identifiables ; c-à-d., si l'application $\theta) \longmapsto P_\theta$ est injective.

On peut affaiblir la notion précédente à une notion locale.

Définition 2.13. *Une valeur du paramètre* $\theta)_0 \in \Theta$ *est localement identifiable s'il existe un voisinage* ω_0 *de* $\theta)_0$ *tel que* $\forall \theta \in \omega_0 : \theta \neq \theta_0$ *on a* $P_\theta \neq P_{\theta 0}$ *. Le modèle* $(\mathbf{X}, P_\theta)$, $\theta \in \Theta$ *est dit localement identifiable si tous les paramètres sont localement identifiables.*

2.3 Éléments de théorie de l'information

On définira dans cette section différentes quantités mesurant l'information contenue dans un modèle statistique.

Information au sens de Fisher

Soit le modèle statistique $(\mathbf{X}, P_\theta)$, $\theta \in \Theta$ tel que P_θ admet une densité $f(x, \theta)$ relativement à la mesure dominante μ. On appellera hypothèses usuelles les 4 hypothèses suivantes :

$H1$: Θ est un ouvert de $\mathbb{R}^d$ pour un certain d fini.

$H2$: Le support $\{x : f(x, \theta) > 0\}$ ne dépend pas de θ.

$H3$: Pour tout $x \in \mathbf{X}$ la fonction $f(x, \theta)$ est au moins deux fois dérivable par rapport à θ pour tout $\theta \in \Theta$ et que les dérivées première et seconde sont continues. On dit que $\theta \longmapsto f(x, \theta)$ est C^2.

$H4$: Pour tout $B \in \mathbf{B}$ l'intégrale $\int\limits_B f(x, \theta) d\mu(x)$ est au moins deux fois dérivable sous le signe d'intégration et on peut permuter intégration et dérivation ; c-à-d.,

$$\frac{\partial}{\partial \theta_j} \int\limits_B f(x, theta) d\mu(x) = \int\limits_B \frac{\partial f(x, \theta)}{\partial \theta_j} d\mu(x), \quad j = 1, ..., d$$

$$\frac{\partial^2}{\partial \theta i \partial \theta_j} \int\limits_B f(x, \theta) d\mu(x) = \int\limits_B \frac{\partial^2 f(x, \theta)}{\partial \theta_i \partial \theta_j} d\mu(x), \quad i, j = \{1, ..., d\}$$

Lorsque ces 4 hypothèses sont vérifiées, on dit que le modèle est régulier.

Exemple 2.14. *Les modèles* $X \rightsquigarrow P(\theta)$, $\theta > 0$, $X \rightsquigarrow Exp(\lambda)$, $\lambda > 0$ *et* $X \rightsquigarrow N(\mu, \sigma 2)$, $\theta = (\mu, \sigma 2) \in \mathbb{R} \times \mathbb{R}_+^*$ *sont réguliers mais pas* $X \rightsquigarrow U[0, \theta]$, $\theta > 0$.

On appelle score le vecteur aléatoire $S(X, \theta)$ définit par

$$S(X, \theta) = \nabla_\theta (log f(X, \theta)) = \left(\frac{\partial log f(X, \theta)}{\partial \theta 1}, ..., \frac{\partial log f(X, \theta)}{\partial \theta_d} \right)^T$$

Propriété – Le score est un vecteur aléatoire centré $E(S(X, \theta)) = 0$.

– Le vecteur score est additif : Soient X et Y deux variables alé atoires indépendantes associées aux modèles statistiques (X, P_θ) et (Y, Q_θ). Alors $S(X,\theta)$ et $S(Y,\theta)$ sont indépendants

$$S((X,Y),\theta) = S(X,\theta) + S(Y,\theta), \forall\theta \in \Theta.$$

Ici (X,Y) est associé au modèle statistique $(X \times Y, P_\theta \otimes Q_\theta)$.

Définition 2.15. *On appelle information de Fisher au point θ la matrice*

$$I(\theta) = E\left(S(X,\theta)S(X,\theta)^T\right) =$$

$$\begin{pmatrix} E\left[\left(\frac{\partial logf(X,\theta)}{\partial\theta_1}\right)^2\right] & E\left[\left(\frac{\partial logf(X,\theta)}{\partial\theta_1}\frac{\partial logf(X,\theta)}{\partial\theta_2}\right)\right] & \cdots & E\left[\left(\frac{\partial logf(X,\theta)}{\partial\theta_1}\frac{\partial logf(X,\theta)}{\partial\theta_d}\right)\right] \\ \vdots & & \vdots & \vdots \\ E\left[\left(\frac{\partial logf(X,\theta)}{\partial\theta_d}\right)^2\right] & E\left[\left(\frac{\partial logf(X,\theta)}{\partial\theta_d}\frac{\partial logf(X,\theta)}{\partial\theta}_2\right)\right] & \cdots & E\left[\left(\frac{\partial logf(X,\theta)}{\partial\theta_d}\right)^2\right] \end{pmatrix}$$

Pour un modèle régulier, on a la relation $I(\theta) = -E\left[\nabla_\theta(S(X,\theta)^T\right] =$

$$\begin{pmatrix} -E\left[\left(\frac{\partial^2 logf(X,\theta)}{\partial\theta_1^2}\right)\right] & -E\left[\left(\frac{\partial^2 logf(X,\theta)}{\partial\theta_1\partial\theta_2}\right)\right] & \cdots & -E\left[\left(\frac{\partial^2 logf(X,\theta)}{\partial\theta_1\partial\theta_d}\right)\right] \\ \vdots & & \vdots & \vdots \\ -E\left[\left(\frac{\partial^2 logf(X,\theta)}{\partial\theta_d^2}\right)^2\right] & E\left[\left(\frac{\partial^2 logf(X,\theta)}{\partial\theta_d\partial\theta_2}\right)\right] & \cdots & E\left[\left(\frac{\partial^2 logf(X,\theta)}{\partial\theta_d^2}\right)\right] \end{pmatrix}$$

et donc pour tout $1 \leq i,j \leq d$: $I_{ij}(\theta) = -E\left[\left(\frac{\partial^2 logf(X,\theta)}{\partial\theta_i\partial\theta_j}\right)\right]$

Notons que pour le calcul de $I(\theta)$, l'espérance est prise par rapport à P_θ, à θ fixé.

Propriété On suppose ici que les hypothèses $H1 - H4$ sont vérifiées, donc que le modèle est régulier.

– L'information de Fischer est une matrice symétrique définie positive. En effet, étant donné que le score est centré $I(\theta) = Var(S(X,\theta)) \geq 0$.

– L'information de Fischer est additive : Si X et Y deux variables aléatoires indépendantes dans des modèles paramétriques au paramètre θ commun alors $I(X,Y)(\theta) = IX(\theta) + IY(\theta)$, $\forall\theta \in, \Theta$

car c'est la variance d'une somme de scores indépendants.

Soit $X \rightsquigarrow N(\mu;\sigma^2)$, alors $I(\mu,\sigma^2)$= $\begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$

En effet,

$$\log f(x,\mu,\sigma^2) = -\frac{1}{2}\log 2\Pi - \Pi 12\log\sigma^2 - \frac{1}{2\sigma^2}(x-)^2$$

$$\frac{\partial^2 logf(x,\mu,\sigma^2)}{\partial\mu^2} = \frac{1}{\sigma^2} \Rightarrow -E\left[\frac{\partial^2 logf(X,\mu,\sigma}{2)}\partial\mu^2\right] = \frac{1}{\sigma^2}$$

$$\frac{\partial^2 logf(x,\mu,\sigma^2)}{(\partial\sigma^2)^2} = \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(x-\mu)^2 \Rightarrow -E\frac{\partial^2 logf(X,\mu,\sigma^2)}{(\partial\sigma^2)^2} = \frac{1}{2\sigma^4}$$

$$\frac{\partial^2 logf(x,\mu,\sigma^2)}{\partial\mu\partial\sigma^2} = 0 \Rightarrow E\frac{\partial^2 logf(X,\mu,\sigma^2)}{\partial\mu\partial\sigma^2 = 0}$$

Pour un échantillon $X_1, ..., X_n$, le vecteur score $S_n(\theta)$ et l'information de Fischer $I_n(\theta)$ associés à θ sont donnés par

$$S_n(\theta) = \nabla_\theta \sum_{i=1}^{n} \log f(X_i,\theta) \quad et \qquad I_n(\theta) = var(S_n(\theta)).$$

On déduit de l'indépendance des X_j que

$$S_n(\theta) = \sum_{j=1}^{n} S(X_j,\theta)$$

où les scores $S(X_1,\theta), ..., S(X_n,\theta)$ sont $(i.i.d)$. (la loi de $S(X,\theta)$ est l'image de la loi de X par l'application $S : x \longmapsto S(x,\theta)$). Etant donné que $E(S(X,\theta)) = 0$, et $Var(S(X,\theta)) = I(\theta) < +\infty$, on a donc la relation $I_n(\theta) = nI(\theta)$.

En vertu de la loi forte des grands nombres et du théorème central limite, on a aussi :

$$\frac{1}{n}S_n(\theta) \to 0p.s \quad et \qquad \frac{S_n(\theta)}{\sqrt{n}}(L) - \to N_d(0, I(\theta))$$

Chapitre 3

ESTIMATION

Objectif : L'estimation consiste à rechercher la valeur numérique d'un ou plusieur paramètres inconnus d'une loi de probabilité à partir d'observations (valeurs prises par la v.a. qui suit cette loi de probabilité)

3.1 Distribution d'échantillonnage

Pour résoudre les problèmes d'estimation de paramètres inconnus, il faut tout d'abord étudier les distributions d'échantillonnage, c'est à dire la loi de probabilité suivie par l'estimateur.

3.1.1 Loi de probabilité de la moyenne

Soit X une variable aléatoire suivant une loi normale d'espérance μ et de variance σ^2 et n copies indépendantes X_1,X_2,...,X_i,...,X_n telle que X_i associe le ième élément de chacun des n échantillons avec $E(X_i) = \mu$ et V(X_i) $=\sigma^2$.

On construit alors la variable aléatoire X , telle que $\overline{X} = \frac{X_1+X_2+..X_i+...+X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$

avec pour espérance : $E(\overline{X}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}E(\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) == \frac{1}{n}\mu$ (Propriétés de l'espérance)

d'où $E(\overline{X}) = \mu$ $E(\overline{X})$ est notée également $\mu_{\overline{X}}$

et pour variance si $V(X_i) = \sigma^2$:

$V(\overline{X}) = V(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}E(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{\sigma^2}{n}$ (Propriétés de la variance)

$V(\overline{X})$ est notée également $\sigma^2_{\overline{X}}$

La loi de probabilité de la variable aléatoire X , moyenne de n v.a. X de loi de probabilité N$(\mu, \sigma), estuneloinormaleN$

Remarque : il est aisé de voir sur un graphe que la variance associée à une moyenne $\frac{\sigma^2}{n}$ est plus faible que

la variance de la variable elle-même (σ^2).

Exemple :

Des études statistiques montrent que le taux de glucose dans le sang est une variable normale X d'espérance $\mu = 1$ g/l et d'écart-type $\sigma = 0{,}1$ g/l.

En prenant un échantillon de 9 individus dans la population, l'espérance et l'écart-type théorique attendu de la variable aléatoire X sont alors :

$\mu_X = \mu = 1g/l$ *et* $\sigma_X = \frac{\sigma}{\sqrt{n9}} = \frac{0,1}{\sqrt{9}} = 0,03g/l$

3.1.2 Convergence

En fonction de la nature de la variable aléatoire continue X, de la taille de l'échantillon n et dela connaissance que nous avons sur le paramètre σ^2, la variable centrée réduite construite avec $\overline{X}$ converge vers différentes lois de probabilité.

Lorsque la variance σ^2 est connue et n grand ($n \geq 30$), on se trouve dans les conditions du théorème central limite et la loi suivie par : $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \to N(0,1)$ loi normale réduite

Ceci reste vrai lorsque n ≤ 30 seulement si la loi suivie par X suit une loi normale.

Lorsque la variance σ^2 est inconnue et X suit une loi normale, la loi suivie par la variable centrée réduite est alors :$\frac{X-\mu}{\sigma}/\sqrt{n} \to T_{n-1}$ loi de student à n-1 degrés de liberté.

Lorsque n ≥ 30, la loi de student tend vers une loi normale réduite (voir convergence).

Lorsque la variance σ^2 est inconnue et X ne suit pas une loi normale, la loi suivie par $\frac{X-\mu}{\widehat{\sigma}/\sqrt{n}}$ n'est pas connue.

3.2 Estimateur

Définition 3.1. *Soient* X_1 , X_2 ,..., X_i , ..., X_n , *n réalisations indépendantes de la variable aléatoire X (discrète ou continue) et* θ *un paramètre associé à la loi de probabilité suivi par X, un estimateur du paramètre* $\theta estunevariablealatoire\Theta$ *fonction des* X_i : $\Theta = f(X_1, X_2, \ldots, X_i, \ldots, X_n)$

Si on considère n observations x_1 , x_2 ,..., x_i , ..., x_n, l'estimateur $\Theta fourniraun eestimationde\theta notegalement\widehat{\theta}$:

$\widehat{\theta} = f(x_1, x_2, \ldots, x_i, \ldots, x_n)$

L'estimation d'un paramètre inconnu, noté θ est fonction des observations résultant d'un échantillonnage aléatoire simple de la population. L'estimateur est donc une nouvelle variable aléatoire construite à partir des données expérimentales et dont la valeur se rapproche du paramètre que l'on cherche à connaître.

L'estimation de θ est une variable aléatoire Θ dont la distribution de probabilité s'appelle la distribution d'échantillonnage du paramètre θ.

L'estimateur Θ admet donc une espérance E(Θ) et une variance V(Θ).

3.2.1 Propriétés

Convergence

L'estimateur Θ doit tendre vers la valeur réelle du paramètre θ lorsque le nombre d'individus étudié augmente. On dit que l'estimateur est convergent.

Si $\forall \epsilon > 0 P(|\Theta - \theta|) > \epsilon) \to 0$ lorsque $n \to \infty$

Ceci équivaut à dire qu'en limite $\Theta \to \Theta$ lorsque $n \to \infty$.

Biais d'un estimateur

Le biais d'un estimateur noté B(Θ) est la différence moyenne entre sa valeur et celle du paramètre qu'il estime. Le biais doit être égal à 0 pour avoir un bon estimateur.

$B(\Theta) = E(\Theta - \theta) = E(\Theta) - E(\theta) = E(\Theta) - \theta = 0$ (voir propriétés de l'espérance)

d'où $E(\Theta) = \theta$

Ainsi l'estimateur sera sans biais si son espérance est égale à la valeur du paramètre de la population $E(\Theta) = \theta$

Remarque : Un estimateur est asymptotiquement sans biais si $E(\Theta) \to \theta$ lorsque n $\longrightarrow \infty$

Variance d'un estimateur

Si deux estimateurs sont convergents et sans biais, le plus efficace est celui qui a la variance la plus faible car ses valeurs sont en moyenne plus proches de la quantité estimée. $V(\Theta) = E(\Theta - E(\Theta))^2$ minimale

Remarque : Quand les estimateurs sont biaisés, en revanche, leur comparaison n'est pas simple. Ainsi un estimateur peu biaisé mais de variance très faible, pourrait même être préféré à un estimateur sans biais mais de grande variance.

Si un estimateur est asymptotiquement sans biais et si sa variance tend vers 0 lorsque $n \longrightarrow \infty$, il est convergent.

$P(|\Theta - \theta| \geq \epsilon) \leq \dfrac{V(\Theta)}{\epsilon^2}$ avec ϵ>0. (Inégalité de Bienaymé-Tchébycheff)

Cette inégalité exprime que si $|\Theta - \theta|$ tend vers 0 quand n augmente, $V(\Theta)$ doit aussi tendre vers 0.

3.2.2 Estimation ponctuelle

L'estimation d'un paramètre quelconque θ est ponctuelle si l'on associe une seule valeur à l'estimateur $\widehat{\theta}$ à partir des données observables sur un échantillon aléatoire.

Si la distribution de la variable aléatoire X est connue, on utilise la m éthode du maximum de vraisemblance pour estimer les paramètres de la loi de probabilité. En revanche si la distribution n'est pas connue, on utilise la méthode des moindres carrés.

3.2.3 Espérance

Soit X une variable aléatoire continue suivant une loi normale $\mathrm{N}(\mu, \sigma) dont la valeur des paramtres n'est pas connue et p$

Soient X_1 , X_2 ,. . . , X_i , . . . , X_n , n réalisations indépendantes de la variable aléatoire X, un estimateur du paramètre $\mu est une suite de variable aleatoire \Theta fonctions des X_i : \Theta = f(X_1, X_2, \ldots, X_i, \ldots, X_n)$

La méthode des moindres carrés consiste à rechercher les coefficients de la combinaison linéaire $\Theta = a_1X_1 + a_2X_2 + \ldots + a_iX_i + \ldots + a_nX_n$ telle que $E(\Theta) = \mu$ et $V(\Theta)$ soit minimale.

La moyenne arithmétique constitue le meilleur estimateur de $\mu, esprance de la loi de probabilit de la variable alatoire X$

$\widehat{u} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Voici pourquoi :

Estimateur sans biais : $E(\overline{X}) = \mu$ (voir loi de la moyenne)

Estimateur convergent : si l'on pose l'inégalité de Biénaymé-Tchébycheff :

$P(\left|\overline{X} - \mu\right| \geq \epsilon) \leq \frac{V(\overline{X})}{\epsilon^2}$ avec ϵ>0

lorsque $n \to \infty$ $\frac{V(\overline{X})}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$ et ceci $\forall\epsilon$>0.

ainsi en limite,$P(\left|\overline{X} - \mu\right| \geq \epsilon) = 0$, ce qui indique que X $\longrightarrow \mu$ en probabilité.

3.2.4 Variance

Soit X une variable aléatoire continue suivant une loi normale $N(\mu, \sigma)$ pour laquelle on souhaite estimer la variance σ^2.

Soient X_1 , X_2 ,. . . , X_i , . . . , X_n , n réalisations indépendantes de la variable aléatoire X, un estimateur du paramètre σ^2 est une suite de variable aléatoire $\Theta fonctions des X_i : \Theta = f(X_1, X_2, \ldots, X_i, \ldots, X_n)$

• Cas où l'espérance μ est connue

La méthode des moindres carrés consiste à rechercher les coefficients de la combinaison linéaire $\Theta = a_1(X_1 - \mu)^2 + a_2(X_2 - \mu)^2 + \ldots + a_i(X_i - \mu)^2 + \ldots + a_n(X_n - \mu)^2$

telle que $E(\Theta) = \sigma^2$ et $V(\Theta)$ soit minimale.

La variance observée constitue le meilleur estimateur de σ^2, variance de la loi de probabilité de la variable aléatoire X lorsque l'espérance μ est connue :

$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$

Remarque : Cette estimation de la variance de la population est rarement utilisée dans la mesure où si la variance σ^2 n'est pas connue, l'espérance μ ne l'est pas non plus.

• Cas où l'espérance μ est inconnue

Dans ce cas, nous allons estimer μ avec $\widehat{\mu} = \overline{X}$ et dans ce cas $\sum_{i=1}^{n}(X_i - \mu)^2 \neq \sum_{i=1}^{n}(X_i - \overline{X})^2$

Nous allons étudier la relation entre ces deux termes à partir de la variance observée :

$s^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(X_i - u) - (\overline{X} - u)\right]^2 = \sigma^2 - \dfrac{\sigma^2}{n}$

en effet $\sigma^2_{\overline{X}} = \dfrac{1}{n}\sum_{i=1}^{n}(\overline{X} - u)^2 = (\overline{X} - u)^2 = \dfrac{\sigma^2}{n}$

ainsi $s^2 = \frac{(n-1)\sigma^2}{n}$

Le meilleur estimateur de σ^2, variance de la loi de probabilité de la variable aléatoire X lorsque l'espérance $\mu estinconnueest : \widehat{\sigma}^2 = \frac{ns^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

Remarque : Lorsque n augmente, la variance observée s^2 tend vers la variance de la population σ^2.

$lim_{n\to\infty} s^2 = lim_{n\to\infty}\frac{(n-1)\sigma^2}{n} = \sigma^2.$

3.2.5 quelques méthodes d'estimation

Les diverses méthodes permettent d'obtenir des estimateurs de qualités différentes

La méthode de maximum de vraisemblance

Définition 3.2. *La statistique* $w \longmapsto \arg max(\theta \longmapsto \prod_{i=1}^{n} f_\theta(X_i(w)))$ *s'appelle l'estimateur de maximum de vraisemblance de* θ.
$L : \theta \longmapsto \prod_{i=1}^{n} f_\theta(x_i)$ *s'appelle la fonction vraisemblance du modèle.*

$l : \theta \longmapsto \sum_{i=1}^{n} \log f_\theta(x_i)$ *s'appelle la fonction log-vraisemblance du modèle.*

En pratique, on fait l'étude de l'une des fonctions L ou l. Il n'y a pas forcément unicité. Ces fonctions ne sont

pas nécessairement dérivables ce qui annule le gradient ne réalise pas forcément un maximum .

Remarque 3. *L'estimateur de maximum de vraisemblance n'existe pas toujours et n'est pas toujours unique.*

Exemple 3.3. *Le modèle de la loi exponentielle*
$\Theta = \mathbb{R}^+,\ f_\theta(x) = \theta e^{-\theta x}$ *on a*

$$L(\theta) = \prod_1^n f_\theta(x) = \theta^n e^{-\theta \sum_{i=1}^n x_i}$$

$$l(\theta) = \log L(\theta) = n \log \theta - \theta \sum_{i=1}^n x_i$$

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^n x_i = 0 \Longleftrightarrow \hat{\theta} = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{X}_n}$$

L est une application concave car on a

$$\frac{\partial^2 L(\theta)}{\partial^2 \theta} = -\frac{n}{\theta}$$

Donc, $\hat{\theta} = \frac{1}{\bar{X}}$ *est l'estimateur de maximum de vraisemblance dans le cas d'un modèle de la loi exponentielle .*

La méthode des moments

L'idée de base est d'estimer une espérance mathématique par une moyenne empirique, une variance par une variance empirique, etc...

Si la loi des X_i a deux paramètres θ_1 et θ_2 tels que $(\mathbb{E}(X), Var(X)) = \varphi(\theta_1, \theta_2)$, où φ est une fonction inversible, alors les estimateurs de θ_1 et θ_2 par la méthode des moments sont : $(\hat{\theta}_{1n}, \hat{\theta}_{2n}) = \varphi^{-1}(\bar{X}_n, S_n^2)$.

Ce principe peut naturellement se généraliser aux moments de tous ordres, centrés ou non centrés : $\mathbb{E}[(X - \mathbb{E}(X))^k]$, et $\mathbb{E}(X^k), k \geq 1$.

Exemple 3.4. *La loi Gamma*
Si $X_1, ..., X_n$ *sont indépendantes et de même loi gamma* $G(\alpha, \lambda),\ \mathbb{E}(X) = \dfrac{\alpha}{\lambda}$ *et* $Var(X) = \dfrac{\alpha^2}{\lambda}$*. On en déduit facilement que*

$$\lambda = \frac{\mathbb{E}(X)}{Var(X)} \text{ et } \alpha = \frac{[\mathbb{E}(X)]^2}{Var(X)}$$

Donc les EMM de α *et* λ *sont*

$$\hat{\lambda}_n = \frac{\bar{X}_n}{S_n^2} \text{ et } \hat{\alpha}_n = \frac{\bar{X}_n^2}{S_n^2}$$

Remarque 4. *Dans certains cas, l'estimation par la méthode des moments est moins bonne que l'estimation par maximum de vraisemblance. Néanmoins, dans le cas de la loi Gamma par exemple, le calcul de la fonction de vraisemblance peut poser des problèmes (l'utilisation de l'ordinateur et d'algorithmes numériques est indispensable) tandis que l'estimation des moments est très facilement accessible.*
Lorsque la taille de l'échantillon n'est pas suffisamment grande, la loi des grands nombres ne s'applique pas et par conséquent, les moments empiriques n'approchent pas suffisamment les moments théoriques.

Chapitre 4

LES TESTS STATISTIQUES

4.1 Introduction

Un test statistique est appelé à dégager un résultat significatif au milieu d'un ensemble de données expérimentales aléatoires.

La méthodologie des tests consiste à répondre à l'aide de résultats expérimentaux à une question concernant les paramétres de la loi de probabilité des variables aléatoires.

Quatre conditions préalables au calcul d'un test doivent être réunies :

– la question doit être posée de telle sorte qu'il n'y ait que deux réponses possibles : oui et non ;

– on doit avoir des données chiffrées résultant d'un échantillon ou d'une expérimentation ;

– ces données doivent pouvoir être considérées comme la réalisation de variables aléatoires dont la forme de la loi de probabilité est connue ;

– la question doit concerner un ou plusieurs paramétres de cette loi.

Une fois posée cette dernière, la réponse du test est :

– soit l'acceptation de l'hypothèse, ce qui signifie que les donneés ne sont pas en contradiction avec l'hypothèse ;

– soit le rejet de cette hypothèse, ce qui signifie qu'il est très peu probable d'obtenir les résultats que l'on a trouvés si l'hypothèse est vraie, ou encore que les données sont en contradiction avec elle.

4.2 La formulation des hypothèses

Un test statistique est un mécanisme qui permet de trancher entre deux hypothèses à partir de résultats observés sur un ou plusieurs échantillons.

Soit H_0 et H_1 ces deux hypothèses. La première appelée **hypothèse nulle**, joue un rôle particulier ; elle prétendra que les différences observées entre valeurs calculées et valeurs

théoriques sont dûes au hasard. Si on doit rejeter l'hypothèse nulle H_0, on dira que les écarts obsevés sont significatifs et on choisira H_1 appelée **hypothèse alternative.** Les tests statistiques permettent de retenir ou de rejeter H_0 qui est la seule hypothèse testée et celle qui permet les calculs pour conduire à la conclusion.

On a

H_0 vraie et H_1 fausse

ou

H_0 fausse et H_1 vraie

Il ya 4 solutions dont seulement les deux premières son justes :

a)-H_0 est vraie et on a choisi H_0

b)-H_0 est fausse et on a rejeté H_0

c)-H_0 est vraie et on a rejeté H_0

d)-H_0 est fausse est on a choisi H_0

4.2.1 Le risque d'erreur

Soit un test qui aboutit à chgoisir H_0 ou H_1. Seule une de ces deux hypothèses est vraie et on peut résumer les différents cas de décision et de validité de cette décision par le tableau suivant :

Hypothèse retenueHypothèse vraie	H_0	H_1
H_0	1-α	β
H_1	α	1-β

De ce tableau on tire les définitions suivantes :

Le risque de première espèce α

On appelle risque de première espèce et on note α la probabilité de rejeter l'hypothèse nulle H_0 alors qu'elle sest vraie.

Dans la pratique des tests statistiques, il est d'usage de choisir α a priori $\alpha = 1\%$ ou 5% dans la plupart des cas, cette probabilité est aussi appelée seuil de signification du test.

Le risque de deuxième espèce

β

On appelle risque de seconde espèce et on note β la probabilité d'accipter l'hypothèse nulle H_0 alors qu'elle sest fausse.

α étant fixé, β est déterminé par un calcul de pribabilité si H_1 est précisément définie.

On appelle puissance du test la probabilité $(1 - \beta)$ de rejeter H_0 en ayant raison.

4.3 Les différents types de tests

1-Les tests de conformité

2-les tests de comparaison

3-Les tests d'ajustement à une loi théorique

4-les tests d'indépendance

4.3.1 Les tests de conformité

1.3.1.1 - Etude des moyennes

-Test bilatéral

Nous nous proposon d'étudier la conformité d'un échantillon par rapport à une norme préalablement définie.

a- position du problème

Dans un laboratoire pharmaceutique, une machine automatique fabrique en grande quantité des suppositoires contenant du paracétamol.

On désigne par X la variable aléatoire, qui à tout suppositoire pris au hasard dans la production, associe la masse (en mg) de paracitamol qu'il contient.

On admet que X suit la loi normale de moyenne m et d'écart-type $\sigma = 0.8$.

On veut contrôler la qualité de fabrication sur une pé riode donnée. Dans ce but, pendant le fonctionnement de la machine, on prélève d'un temps à l'autre un suppositoire dont on mesure la masse du paracétamole. On constitue ainsi un échantillon de 100 suppositoires. Les tirages sont supposéss indépendants.

On se propose de construire un test bilatéral permettant d'accepter ou de refuser, au seuil de signification de 5%, l'hypothèse selon laquelle la masse moyenne de paracétamol contenue dans un suppositoire est égale à 170 mg.

l'hypothèse nulle H_{00} est m=170 mg et lhypothèse alternative est H_1 $m \neq 170$ mg.

1- Sou H_0 quelle est la loi de la variable alétoire $\overline{X}$? préciser ces paramètres.

2- Enoncer clairement la règle de décision du test

3- Les résultats des mesures de l'échantion prélevé sont donnés dans le tableau :

Masse (mg)	$[145;155[$	$[155;165[$	$[165;175[$	$[175;185[$	$[185;195[$
Effectifs	7	30	43	16	4

Peut-on accepter l'hypothèse H_0 au seuil de signification de 5% ?.

b - Lois d'échantillonnage

Puisque $n \geq 30$, le théhrème de la limite centrée nous permet de dire que la variable aléatoire $\overline{X}$ qui à chaque échantillon de taille n associe sa moyenne, suit approximativement la loi normale $N(u; \frac{\sigma}{\sqrt{n}})$. Alors la variable aléatoire $T = \dfrac{\overline{X} - u}{\dfrac{\sigma}{\sqrt{n}}}$ suit la loi normale centrée réduite.

c- Construction d'un test bilatéral :

l'hypothèse nulle $(H)_0$ est m=170 mg et l'hypothèse alternative est (H_1) m$\neq$ 170mg

e- Règle de décision :

Fixon, 'a priori, le risque maximal que nous acceptons de prendre en refusant H_0 alors qu'elle est vraie. Ce risque dit de première espèce, et noté α.

Puisque T suit la loi normale centrée réduite, il existe un unique réel strictyement positif t_α tel que :$p\,|T| > t_\alpha = \alpha.t_\alpha = \Pi^{-1}\left(1 - \frac{\alpha}{2}\right)$

Si $|T| > t_\alpha$ on rejette H_0 avec le risque α de se tromper.

Si $|T| \leq t_\alpha$ on accepte H_0 avec le risque de se tromper.(risque β de seconde espèce non quantifié).

Application numérique

Sous l'hypothèse H_0 la variable aléatoire $\overline{X}$ suit la loi normale $N(170; 0,8)$ donc la variable aléatoire $T = \frac{\overline{X} - 170}{0,8}$ suit la loi normale centrée réduite.

au seuil de risque $\alpha = 0,05$ on rejette H_0E si $|T| > 1,96$.

Pour l'échantillon proposé, en utilisant les centres des classes, on trouve $\overline{x}$= 168

On en déduit $t = -2,5$ donc $|t| > 1,96$ et on rejette H_0 au risque de 5% de se tromper.

4.3.2 les tests d'homogénéité (grands échantillons)

Nous disposons de deux échantillons indépendants donnés sous la forme d'un tableau d'éffectifs ou de fréquences du caractère étudié.

Nous désirons savoir si les différences observées sur la moyenne ou sur le fréquence sont dues u'niquement au hasard de l'échantillonage ou si elle sont trop importantes est doivent être atribuées à d'autres causes.

1.3.2.1.Etude des moyennes

a - Position du problème On étudie ici un caractère quantitatif C et on dispose de deux grands échantillons indépendants

A d'effectif n_A , de moyenne m_A et d'écrt-type σ_A

B d'effectif n_B , de moyenne m_B et d'écrt-type σ_B

A quelles condition peut-on conclure, qu'à un risque doné, ces deux échantillons proviennent de la même population ?

b - Lois d'échantillonnage Supposons que l'échantillon A provienne de la population P , d'effectif N, de moyenne u et d'écart-type σ.

Supposons que l'échantillon B provienne de la population P$'$, d'effectif N$'$, de moyenne u$'$ et d'écart-type σ'.

On sait que si $N_A \geq 30$, La variable aléatoire $\overline{X}$ qui à tout échantillon de taiile n_A associe sa moyenne m_A suit apprximativement la loi normale $N(u; \frac{\sigma}{\sqrt{n_A}})$.

Même si $N_B \geq 30$, La variable aléatoire $\overline{X}$ qui à tout échantillon de taiile n_B associe sa moyenne m_B suit apprximativement la loi normale $N(u'; \frac{\sigma'}{\sqrt{n_B}})$

Les variables alétoires $\overline{X}_A$ et $\overline{X}_B$ etant indépendantes et La variable aléatoire $\overline{X}_A - \overline{X}_B$ suit approximativement la loi normale $N(u-u'; \sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma'^2}{n_B}})$.

Tests d/hypothèse bilatéral

a - Hypotèse à tester Nous nous proposons de tester l'hypothèse nulle, notée H_0 "u et u' ne sont pas significativement différentes"

b - Hypothèse alternative H_1 : le test étant bilateral H_1 est u et u' sont significativement différentes"

c - Règle de dcision : Sous l'hypothèse H_0, la variable alétoire $\overline{X}_A - \overline{X}_B$ suit approximativement la loi normale $N(0; \sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma'^2}{n_B}})$.

Donc la variable alétoire $\dfrac{\overline{X}_A - \overline{X}_B}{\sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma'^2}{n_B}}}$ suit approximativement la loi normale $N(0;1)$.

Fixons alors un seuil de risque α (donc un seil de confiance $1-\alpha$), on sait qu'il existe un réel unique t_α strictement positif tel que $P(|T| \leq t_\alpha) = 1-\alpha$

$P(|T| \leq t_\alpha) = 1-\alpha$ é quivaut à $t_\alpha = \Pi^{-1}\left(1 - \dfrac{\alpha}{2}\right)$

La règle de décision du test en résulte :

Si $|T| \leq t_\alpha$ on a aucune raison de rejeter H_0 donc on l'accepte.avec un risque β (non contifié) de se tromper

Si $|T| > t_\alpha$ on rejette H_0 un risque α de se tromper

d - Mise on oeuvre du test : $t = \dfrac{m_A - m_B}{\sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma'^2}{n_B}}}$

On compare alors $|t|$ avec t_α et on utilise la règle de décision pur conclure.

En général σ et σ' sont inconnus et remplacés dans cette formule par $\widehat{\sigma}_A = \sigma_A\sqrt{\frac{n_A}{n_A-1}}$ et $\widehat{\sigma}_B = \sigma_A\sqrt{\frac{n_B}{n_B-1}}$

Définition 4.1. *Soient X et Y deux variables aléatoires indépendantes suivant respectivement $N(0,1)$ et $X^2(n)$. On appelle loi de Student à n degrés de liberté la loi suivie par le rapport : $T = \frac{X}{\sqrt{Y/n}}$, cette loi est notée T_n.*

$E(T_n) = 0 \quad (n > 1)\,; \;\; Var(T_n) = \dfrac{n}{n-2} \qquad (n > 2).$

Définition 4.2. *Soient X et Y deux variables aléatoires indépendantes suivant respectivement $X^2(n)$ et $X^2(m)$.*

La variable aléatoire $F = \frac{EX/n}{EY/m}$ suit la loi de Fisher-Snedecor à n et m degrés de liberté notée $F_{n,m}$

$E(F_{n,m}) = \dfrac{1}{m-2} \qquad (m > 2); \;\; Var(T_n) = \dfrac{2m^2(n+m-2)}{n\,(m-4)\,(m-2)^2} \qquad (m > 4).$

4.3.3 Test de Student

On pratique il est rare que l'on connaisse la valeur de σ ; on n'en connait q'une estimation s, valeur calculée de l'estimateur S. Que peut-on dire alors de la variable centrée réduite $\dfrac{\overline{X} - m}{S/\sqrt{n}}$?.

Sous réserve que le caractère étudié soit distribué dans la population selon la loi normale, on peut démontrer que ce rapport suit une loi de Student 0 $(n-1)$ degré de liberté et que cette loi converge rapidement vers la loi de Gauss lorsque n augmente, peut être remplacée par elle dès que $n \geq 30$.

On voit donc que pour les petits échantillons ($n < 30$), il faut faire appel à la loi de Student.La comparaison de moyennes à partir de petits échantillons (n_1 et / ou $n_2 < 30$) va elle aussi utiliser cette loi de Student.

Faisons l'hypothèse que les deux échantillons proviennent de populations de mêmes moyennes (il s'agit de l'hypothèse $m_1 = m_2 = m$) et qu'en outre ses populations sont normales et de mêmes variances($\sigma_1^2 = \sigma_2^2 = \sigma^2$), on peut démontrer que la quantité $t = \dfrac{|\overline{x}_1 - \overline{x}_2|}{\widehat{\sigma}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

où $\begin{cases} \widehat{\sigma}^2 = \dfrac{n_1\sigma_{e1}^2 + n_2\sigma_{e2}^2}{n_1 + n_2 - 2} \\ \overline{x}_i,\ \sigma_{ei}^2 \text{ moyenne et écart-type de l'échantillon numéro i} \end{cases}$ suit la loi de Student à $v = n_1 + n_2 - 2$ degrés de liberté. Il devient alors possible de déterminer une région d'acceptation de l'hypothèse nulle H_0 d'égalité des moyennes. Cette région dépend de l'hypothèse alternative H_1, dans le cas où H_1 est "$m_1 \neq m_2$", on mène un test bilatéral et la région d'acceptation de H_0 est donnée par l'ntervalle : $\left]-t_{v;\alpha}; + - t_{v;\alpha}\right]$, avec $v = n_1 + n_2 - 2$. où $t_{v;\alpha}$ désigne la valeur de la loi de Student ayant la probabilité α d'être dépassée en valeur absolue.

Si $t < t_{v;\alpha}$ alors on accepte H_0.

Si $t > t_{v;\alpha}$ on rejette H_0 au seuil $\alpha\%$.

Remarque : Dans le cas d'une hypothèse alternative conduisant à mener un test unilatéral du type H_0 : "$m_1 = m_2$" contre H_1 : "$m_1 > m_2$" la région d'acceptation de H_0 est de la forme : $]-\infty; t_{2\alpha;v}[$.

.On sera souvent amené à tester de façon préalable l'égalité des variance à l'aide d'un test de Fisher-Snedecor avant de comparer les moyennes à ppartir de deux petits échantillons.

4.3.4 Test de Fisher-Snedecor

Comparaisons de deux variances

Le test de comparaison de deux variances σ_1^2 et σ_2^2 est basé sur le rapport des deux estimation s_1^2 et s_1^2 calculées à partir d'échantillons, de taille respective n_1 et n_2 extraits des deux population à comparer. Il n'est pas nécessaire que n_1 et n_2 soient grand mais il est impératif que les deux populations soient normalement distribuées.

On formule l'hypothèse H_0 : "$\sigma_1^2 = \sigma_2^2$" contre l'hypothèse H_1 : "$\sigma_1^2 \neq \sigma_2^2$" ce test est donc toujours bilatéral. On calcule la quantité : $F = \frac{s_1^2}{s_1^2}$ si $s_1^2 \geq s_2^2$. F est toujours supérieure ou égale à un.

La règle de décision est la suivante :

Si $F < F_{\frac{\alpha}{2};v_1;v2}$ on accepte H_0.

Si $F \geq F_{\frac{\alpha}{2};v_1;v2}$ on rejette H_0. au risque α.

4.4 Le test chi-deux

4.4.1 INTRODUCTION

Problème 1 :

Partant des races pures, un sélectionneur a croisé de mufliers ivoires avec des mufliers rouges, il a obtenu en $F1$ des mufliers pâles, puis en $F2$ après autofécondation des plantes de la génération F1 : 22 mufliers rouges, 52 mufliers pâles et 23 mufliers ivoires.

La couleur des fleurs est-elle gérée par un couple d'allèles ?.

Le test chi-deux est fréquemment utilisé par les biologistes. A la différence des autres test, ce test ne s'appuie pas sur un modèle probabiliste rigoureux, mais sur une loi asymptotique ; il est donc délicat à utiliser et il est parfois préférable de le remplacer, lorsque c'est possible, par un test non parmétrique plus adapté.

Le test du χ^2 est le plus célèbre des tests dits **non paramètriques** qui n'exigent aucune condition sur la distriution de la population mère. C'est un test globale qui porte sur l'ensemble des effectifs ou fréquences observées après expérience et calculés à partir de l'hypothèse testée. On pourra comparer :

Une distribution expérimentale et une distribution théorique. Les caractéristiques de cette distribution théorique sont connues ou estimées à partir des observations. Selon le cas, on parlera de test de **conformité ou d'ajustement** à une loi théorique.

Plusieurs distributions pour savoir si on peut accépter l'hypothèse qu'elles proviennent de la même population parente, dans ce cas on mènera un test **d'homogéneité** ou

d'indépendance. On a en fait généraliser le cas précédent en comparant chaque distribution empirique à une même distribution théorique.

Le mécanisme du test du χ^2 permet de savoir si les écarts constatés entre les distributions à comparer sont imputables ou non au hasard.

Définition 4.3. *Soit X une v.a de loi.N(0 ;1), alors la v.a X^2 est dite v.a de chi-deux à 1 degé de liderté.*

Définition 4.4. *Soient X_1, X_2,...,X_n n v.a indépendantes suivent toutes loi N(0 ;1), alors la v.a $Z = X_1^2 + X_2^2 + ... + X_n^2$ est une v.a de chi-deux à n degés de liderté, avec E(Z)=n et Var(Z)=2n*

Remarque 5. *Si Z suit la loi du χ^2 à n degrés de liberté, la table du chi-deux donne pour un risque α choisi, le nombre χ^2_α tel que*

$$P(Z \geqslant \chi^2_\alpha) = \alpha.$$

4.4.2 COMPARAISON ET AJUSTEMENT A UNE LOI THEORIQUE

Construction du test

On considère une distribution expérimentale donnée par un échantillon de taille n.

Les individus de cet échantillon son classés et on a dénombré la fréquence absolue ou effectif de chaque classe. On note n_i l'effectif observé pour la classe N° i. Si on connait (ou croit connaître) la loi théorique que suit cette distribution, on est alors capable de calculer les effectifs théoriques de chaque classe. En effet la loi théorique est connue dès lors que les probabilités attachées à chaque classe le sont. On note P_i la probabilité qu'un individu tiré au hasard appartienne à la classe N° i. L'effectif théorique associé est alors nP_i.

4.4.3 Application du test chi-deux

On expliquera d'abord les principes du test sur une loi multinomiale puis dans ses applications les plus courantes, la méthode non paramétrique qui en découle.

<u>test sur une loi multinomiale</u>

Distribution à deux classes.

Soit une expérience aléatoire E susceptible

d'entraîner la réalisation d'un évenement E_1 de probabilité $P(E_1)$, ou d'un évenement E_2 de probabilité $P(E_2)$, E_1 et E_2 formant un système complet c-à-d $P(E_1)+P(E_2)=1$ et $P(E_1\cap E_2)=0$.

Soit un ensemble de n expériences identiques à E et indépendantes. On lui associe les variables X_1 et X_2 représentant respectivement le nombre d'événement de E_1 et de E_2 que l'on peut observer $(X_1+X_2=n)$, la réalisatin effective des n expérience entraîne

l'observation des valeurs x_1 de X_1 et x_2 de X_2 $(x_1+x_2=n)$, On dit que les résultats sont reparties en deux classes. On désire tester l'hypothèse H_0 "$P(E_1)=P_1$ et $P(E_2)=P_2$ contre l'hypothèse H_1

"$P(E_1)\neq P_1$ et $P(E_2)\neq P_2$".

Compte-tenue de la relation $P_1+P_2=1$, il suffit de tester "$P(E_1)=P_1$" contre

"$P(E_1)\neq P_1$". Ce que l'on peut faire à l'aide de la variable $X_1\rightarrow B(n,P_1)$.

X_1 admet pour loi asymptotique, lorsque n augmente indéfiniment, la loi

$$N(nP_1,nP_1(1-P_1)).$$

Alors un test avec la variable

$Y=\dfrac{X_1-nP_1}{\sqrt{nP_1(1-P_1)}}$ considéré comme pratiquement normale centrée et réduite sou H_0.

Soit maintenant la variable

$Z=\dfrac{(X_1-nP_1)^2}{nP_1}+\dfrac{(X_2-nP_2)^2}{nP_2}$,

on a $Z=\dfrac{(X_1-nP_1)^2}{nP_1(1-P_1)}=Y^2$

étant donné le comportement asymptotique de Y, il est clair que Z admet pour loi asymptotique la loi de χ_1^2 sous H_0.

Pour un niveau α on peut ecrire $1-\alpha=P\left(-y_{\frac{\alpha}{2}}\leq Y\leq y_{\frac{\alpha}{2}}\right)=P\left(0\leq Y^2\leq y_{\frac{\alpha}{2}}^2\right)=$

$P\left(0\leq Z\leq z_{\frac{\alpha}{2}}\right)$ avec $z_{\frac{\alpha}{2}}=y_{\frac{\alpha}{2}}^2$,

La borne supérieur de l'intervalle d'acceptation $(3.481=(1.96)^2$ au niveau $5\%;6.635=(2.576)^2$ au niveau $1\%)$ étant lue dans les tables de χ^2..

Distribution à r classes.

Plus généralement soit une expérience aléatoire E susceptible d'entraîner la réalisation de r évenements E_1,E_2, ...,E_r de probabilité $P(E_1)$, $P(E_2), ..., P(E_r)$, E_1 E_2 $...E_r$,formant un système complet c-à-d $P(E_1) + P(E_2) + ... + P(E_r) = 1$ et $P(E_i \cap E_j) = 0$ *pour* $i \neq j$.

Les résultats de n expériences identiques à E et indépendantes sont donc réparties en r classes. A un tel ensemble d'expériences, On associe les variables X_1 X_2 $...X_r$ représentant respectivement les effectifs des classes que l'on peut observer,

Le système $(X_1, X_2, ..., X_r)$, forme un sytème multinomial, on veut tester l'hypothèse

"$p(E_1) = p_1$ et $p(E_2) = p_2$ *et...* $p(E_r) = p_r$"

contre l'hypothèse H_1 :

"$P(E_1) \neq P_1$ ou $P(E_2) \neq P_2$...ou $P(E_r) \neq P_r$"

En fait il n'y a parmi r variables que $(r-1)$ variables indépendantes ; En effet les variables sont liées par la relation $X_1 + X_2 + ... + X_r = n$, dès que le hasard attribué une valeur numérique à r-1 variables, la valeur de la dernière est imposée.

APPLICATION :

Problème 1 (solutions) :

Solution : Soient p_1, p_2, p_3 les probabilités pour q'une plante de la génération F2 ait respectivement des fleurs rouges, pâles ou ivoires, soient X_1, X_2 et X_3 les variables représentant les plantes à fleurs rouges, pâles ou ivoires que l'on peut observer sur 97 plantes.

On est amené à tester, après un raisonnement génétique élémentaire, l'hypothèse H_0 :

$p_1 = \frac{1}{4}$, $p_2 = \frac{1}{2}$, $p_3 = \frac{1}{4}$ contre l'hypothèse $H_1 : p_1 \neq \frac{1}{4}$ *ou* $p_2 \neq \frac{1}{2}$ *ou* $p_3 \neq \frac{1}{4}$.

D$'$où le tableau :

phénotypes	rouge	pâle	ivoir	total
probabilité	1/4	1/2	1/4	1
effectif théorique	24.25	48.5	24.25	97
effectif observé	22	52	23	97

-Les conditions d'application de χ^2 sont s'atisfaites, à savoir :

-Les classes constituent un systèmes complet

d'événements ;

-Les 97 expériences sont identiques et
indépendantes ;

-Leur nombre est assez élvé ;

-Les effectifs théoriques sont suffisament
élevés.

Dans ces conditions, sous H_0, la variable $Z = \sum_{i=1}^{3} \frac{(X_i - 97p_i)^2}{97p_i}$ est pratiquement une variable X^2, on effectue un test. L'intervalle d'acceptation de H_0 est, au niveau

5% : $[0\ ;\ 5,991]$.

On a observé la valeur $Z_0 = \frac{(2,25)^2}{24.25} + \frac{(3,50)^2}{48,5} + \frac{(1,25)^2}{24,25} \simeq 0.52.$

Conclusion :

Au niveau 5% on peut accepter l'hypothèse que La couleur des fleurs est gérée par un couple d'allèles.

4.4.4 Tests d/hmogénéité

Principe

Le test χ^2 est également utilisé pour la comparaison de plusieurs échantillons. Le principe du test va être exposé dans un exemple à deux échantillons. on le généralise sans peine pour plusieurs échantillons.

Problème 2 :

On a étudié sur deux échantillons prevenant de deux populations différentes la répartition des quatre groupes sanguins : O, A, B,AB les résultats obtenus sont réparties dans un tableau dit tableau de contingence, à deux lignes et à quatre colonnes :

Groupe	O	A	B	AB	tot
1^{er} éch	121	120	79	33	353
2^{em} éch	118	95	121	30	364
total	239	215	200	63	717

On veut tester l'hypotèse H_0 " les quatre groupes sanguins sont réparties de la même manière sur les deux populations"

contre l'hypothèse H_1 "les répartitions sont différentes".

Sous H_0. la probabilité, pour un individu prélevé au hasard, d'être d'un groupe donné est la même dans les deux populations, on ne connaît pas cette probabilité, sinon le problème serait résolu ; on peut cependant l'estimer et, toujours sous H_0. La meilleure estimation que l'on puise en donner est la proportion des indiviudus de ce groupe observée sur l'ensemble des deux échantillons. C'est ainsi que l'on obtient les estimations :

Pour le groupe O	$p_1 = 239/717 \simeq 0,333$
Pour le groupe A	$p_2 = 215/717 \simeq 0,300$
Pour le groupe B	$p_3 = 200/717 \simeq 0,249$
Pour le groupe AB	$p_4 = 63/717 \simeq 0,088$

$p_1 + p_2 + p_3 + p_4 = 1$. La relation $p_1 + p_2 + p_3 + p_4 = 1$ montre qu'en fait il suffit de trois paramètres pour déterminer complètement le modèle. On déduit de l'estimation précédente les éffectifs théoriques de chaque classe pour un échantillon de taille 353 d'une part et pour un échantillon de taille 364 d'autre part. D'où le tableau :

Groupe	O	A	B	AB	total
1^{er} éch	121 (117,7)	120 (105,9)	79 (98,5)	33 (31)	353
2^{em} éch	118 (121,3)	95 (109,1)	121 (101,5)	30 (32)	364
total	239	215	200	63	717

les effectifs théhoriques sont entre parenthèses, on a par exemple, 117=0,333.353.

Soient maintenant les variables X_1, X_2, X_3, X_4 représentant les effctifs des classes du premiers échantillon et Y_1, Y_2, Y_3, Y_4 représentant les effctifs des classes du deuxième échantillon.

On pose :

$$Z = \frac{(X_1 - 117,7)^2}{117,7} + \frac{(X_2 - 105,9)^2}{105,9} +$$
$$\frac{(X_3 - 98,5)^2}{98,5} + \frac{(X_4 - 31,0)^2}{31,0} + \frac{(Y_1 - 121,3)^2}{121,3} +$$
$$\frac{(Y_1 - 109,1)^2}{109,1} + \frac{(Y_1 - 101,5)^2}{101,5} + \frac{(Y_4 - 32,0)^2}{32}.$$

Les conditions d'application du test χ^2 étant satisfaites pour chaque échantillon, sous H_0, la variable Z peut être considérée comme la somme de deux variables χ^2, l'indépendance des deux séries d'observations permet de considérer la varible Z comme une variable χ^2. On est tenté de dire qu'il s'agit d'une variable χ^2

à $2(4-1)=6$ degrés de liberté ; cependant, l'estimation, à partir des observations des trois paramètres qui déterminent coplètement le modèle probabiliste baisse le nombre de degrés de liberté de 6 à 3. D'où $Z \to \chi^2_3$.

Les valeurs élevées de Z étant plus probables sous H_1 que sous H_0.

Au niveau 5% l'intervalle d'acceptation est $[0;\ 7,815]$, et comme $Z \simeq 11.74 > 7,815$ donc

on peut conclure au rejet de H_0.

C'st-à-dire les quatre groupe sanguins sont réparties différemment sur les deux populations d'où proviennent les deux échantillons. Même au niveau 1% on rejetterait H_0.