Projet tutoré
01/05/21-01/06/21

Présenté par :

MAGHOUS Abdellah

Pour obtenir le diplôme de

Licence fondamentale : Statistiques et Applications

Sur le test de chi-2
application en biologie

soutenue le : *3 Juin 2021*

tuteur et enseignant-référent :*Sandrine DALLAPORTA*

Année universitaire : 2020/2021

Remerciement

Je tiens à remercier Mme Dallaporta de m'avoir accepté pour ce stage, c'est un honneur de travailler sous votre encadrement. Vos conseils m'ont guidé, et vous m'avez fourni tous les moyens nécessaires à l'élaboration de ce travail.

Je tiens à vous exprimer ma reconnaissance et ma gratitude pour la disponibilité, le temps que vous m'avez consacré, l'amabilité et la générosité dont vous avez fait preuve.

Veuillez accepter, mes sentiments les plus respectueux et mes vifs remerciements.

Bref descriptif du LMA Le Laboratoire de Mathématiques de l'Université de Poitiers regroupe l'ensemble des chercheurs en mathématiques de l'université de Poitiers. Il est composé de quatre équipes. Il a pour double objectif de développer une recherche fondamentale en mathématiques et de promouvoir une recherche appliquée en privilégiant autant que possible des interactions avec les sciences expérimentales. Pour mon stage je l'ai passé au sein de l'équipe probabilités, statistique et application .

résumé La statistique est une discipline des mathématiques qui nous permet de collecter des données, de les traiter, de les interpréter afin de les rendre les plus compréhensibles possibles pour tous. Grâce à la statistique on peut effectuer des tests et des simulations sur un échantillon de données d'une population dans différents domaines tels que : le trafic urbain, la gestion d'un hôpital, l'évolution d'une population, les prévisions du cours etc. Et cela ce fait en utilisant les lois de probabilités mathématiques discrètes tel que la loi de Bernoulli, loi Multinomiale,... ou continues telles que la loi Normale, la loi de Student, la loi Fisher et la loi de khi deux qui est l'objet de ce stage. Nous proposons de faire un test statistique aves les test d'hypothèses en utilisant la loi de χ^2 car, le test du χ^2 fournit une méthode pour déterminer la nature d'une répartition, qui peut être continue ou discrète.
Mots-clés : La Statistique, Probabilité, Test d'hypothèse, Loi de Khi deux.

Abstract Statistics is a mathematical discipline that allows us to collect data, processed from the interpreted to the most comprehensible possible for all. The statistic can perform tests and simulations on a data sample of a population in different field such as : urban traffic, the management of a hospital, the evolution of a population, etc. And that this fact using discrete mathematics probability distributions as Bernoulli's law, the Multinomial law ..., or continuous as the Normal Law, Law Student and chi-squared law that is the subject of this training period, we propose to do a statistical test assumptions with the test using the law of chi-squared because the Chi-squared test provides a method for determining the nature of a law, which can be continuous or discrete.
Keywords : Statistics, Probability, Test a hypothesis test , Chi-Squared law.

Table des matières

Introduction

Le test du χ^2 , prononcer « khi-deux » ou « khi carré », est un test statistique permettant de tester l'adéquation d'une série de données à une famille de lois de probabilités ou de tester l'indépendance entre deux variables aléatoires.

Ce test permet de vérifier si un échantillon d'une variable aléatoire X donne des observations comparables à celles d'une loi de probabilité $\mathbb{P}$ définie a priori dont on pense, pour des raisons théoriques ou pratiques, qu'elle devrait être la loi de X. L'hypothèse nulle (H_0) d'un test du χ^2 d'adéquation (dénommé aussi test du χ^2 de conformité ou test du χ^2 d'ajustement) est donc la suivante : la variable aléatoire Y suit la loi de probabilité $\mathbb{P}$.

En termes de p-valeur, l'hypothèse nulle (l'observation est suffisamment proche de la théorie) est généralement rejetée lorsque $p \leq 0.05$.

Le test du χ^2 est un test statistique non paramétrique qui convient à des fréquences donc à des proportions et à des probabilités. Un phénomène quelconque peut être mesuré selon sa fréquence d'occurrence, sa durée, son intensité ou en fonction d'autres caractéristiques.

Motivation J'ai choisi ce sujet pour approfondir mes connaissances dans la théorie du test et applications notamment sur le test de khi-deux car je vais faire un master en statistique appliquée qui s'appuie sur les thèmes probabilité, estimation et test et me diriger vers une carrière de data scientist.

Chapitre 1

Introduction aux tests statistiques

1.1 Test d'hypothèse

En statistiques, un test d'hypothèse est une démarche consistant à rejeter ou à ne pas rejeter (rarement accepter) une hypothèse statistique, appelée hypothèse nulle, en fonction d'un jeu de données (échantillon). Il s'agit de statistique inférentielle : à partir de calculs réalisés sur des données observées, nous émettons des conclusions sur la population, en leur rattachant des risques de se tromper.

On va travailler dans un modèle $(\Omega, \mathfrak{F}, (P_\theta)_{\theta \in \mathfrak{H}})$ avec une suite $X_1, X_2....X_n$ de variables aléatoires indépendantes et identiquement distribuées (iid) sur $(\Omega, \mathfrak{F})$.Le principe général d'un test d'hypothèse est d'étudier une population dont les éléments possédant un caractère mesurable ou qualitatif et dont la valeur du paramétre relative étudié est inconue .

Définition 1.1. *Un test est une fonction mesurable $\phi : \mathfrak{F} \longmapsto [0, 1]$, on refuse l'hypothèse H_0 lorsque $\phi(X) = 1$ et on ne la rejette pas lorsque $\phi(X) = 0$.*

Définition 1.2. *Une statistique de test est une fonction des observations $X_1, X_2....X_n$ qui ne dépend pas du paramètre inconnu θ.*

Construire un test de l'hypothèse nulle H0 contre l'hypothèse alternative H1, c'est établir un critére de décision permettant de choisir entre l'hypothèse H0 et H1. Pour cela, il faut :

1. Préciser les deux hypothéses H_0 et H_1.
   par exemple $H_0 : \theta \in \mathbb{H}_0$ contre$H_1 : \theta \in \mathbb{H}_1$ avec $\mathbb{H}_0 \cap \mathbb{H}_1 = \emptyset$
2. Déterminer la loi du statistique sous H_0 et sous H_1.
3. Choisir le niveau α du test pour déduire les domaines d'acceptation et de rejet de H_0.

Quatre situations doivent être envisagées :

1. L'acceptation de l'hypothèse nulle alors qu'elle est vraie.
2. Le rejet de l'hypothèse nulle alors qu'elle est vraie.
3. L'acceptation de l'hypothèse nulle alors qu'elle est fausse.
4. Le rejet de l'hypothèse nulle alors qu'elle est fausse.

Définition 1.3. *On appelle région de rejet H_0 l'ensemble $\{\omega \in \Omega | \phi(\omega) = 1\}$*

Définition 1.4. *Le test est de niveau $\alpha \in [0,1]$ ssi $\mathbb{P}_\theta(\phi = 0) \geq \alpha$ pour tout $\theta \in \mathbb{H}_0$*

1.1.1 Test du rapport de vraisemblance

cas de 2 hypothèses simples

Définition 1.5. *H_0 (resp. H_1) est appelée hypothèse simple si l'ensemble $\mathbb{H}_0$ (resp.$\mathbb{H}_1$) est réduit à un seul point. exemple $H_0 : \theta = \theta_0$. Elle est dite composite dans le cas contraire.*

on teste $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$

Pour construire un test sur les paramètres d'une loi on utilise le rapport de vraisemblance.

1. si $X_1, X_2....X_n$ échantillon de loi discrète (P_θ) $L(X_1, X_2....X_n) = \prod_{i=1}^{n} P_\theta(X_i = x_i)$
2. si $X_1, X_2....X_n$ échantillon de loi de densité (f_θ) $L(X_1, X_2....X_n) = \prod_{i=1}^{n} f_\theta(x_i)$

Définition 1.6. *L'estimateur du max de vraisemblance est la valeur si elle existe du paramétre θ sous laquelle ce qu'on a observé est le plus probable*

$$\hat{\theta} = argmax\ L(X_1, ..., X_n, \theta)\ \ \theta \in \mathfrak{H}$$

Rapport de vraisemblance V le rapport de vraisemblance s'écrit de la façon suivante

$$V(X_1, X_2....X_n) = \frac{L(X_1, X_2....X_n, \theta_1)}{L(X_1, X_2....X_n, \theta_0)}$$

Par convention :

$V = 0$ si $L(X_1, X_2....X_n, \theta_0) = 0$ et $L(X_1, X_2....X_n, \theta_1) = 0$

$V = \infty$ si $L(X_1, X_2....X_n, \theta_0) = 0$ et $L(X_1, X_2....X_n, \theta_1) > 0$

Sous $H_1 : \theta = \theta_1$ V a tendance à être grand

Sous $H_0 : \theta = \theta_0$ V a tendance à être petit

Remarque 1.7. *on teste $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$ au niveau α est le test de zone de rejet $R = \{V \geq v_\alpha\}$ où $\mathbb{P}_{\theta_0}(R) \leq 0.05$*

Exemple 1.8. *On fait un test pour savoir si la pièce est truquée ou non pour cela on la lance 10 fois. poson $X_i = \mathtt{1}_{pile\ ou\ i-ème\ lancer}$ où les $X_1, ...X_{10}$ sont indépendantes et identiquement distrubuées(iid),qui*

suivent une loi de bernoulli de paramètre θ avec

$$L(X_1, X_2....X_n, \theta) = \prod_{i=1}^{10} \theta^{X_i}(1-\theta)^{1-X_i}$$

donc

$$L(X_1, X_2....X_n, \theta) = \theta^{\sum_{i=1}^{10} X_i}(1-\theta)^{10-\sum_{i=1}^{10} X_i} = \theta^{T_{10}}(1-\theta)^{1-T_{10}}$$

on test $H_0 : \theta = \theta_0 = 0,3$ contre $H_1 : \theta = \theta_1 = 0,5$.
on calcule le test du rapport de vraisembalance :

$$V(X_1, X_2....X_{10}) = \frac{L(X_1, X_2....X_n, \theta_1)}{L(X_1, X_2....X_n, \theta_0)} = \frac{\theta_1^{T_{10}}(1-\theta_1)^{1-T_{10}}}{\theta_0^{T_{10}}(1-\theta_0)^{1-T_{10}}}$$

la zone de rejet est

$$\begin{aligned} R = \{V \geq v_\alpha\} &= \{(\frac{\theta_1(1-\theta_1)}{\theta_0(1-\theta_0)})^{T_{10}} \geq v_\alpha\} \\ &= \{T_{10}\log(\frac{\theta_1(1-\theta_1)}{\theta_0(1-\theta_0)}) \geq \log(v_\alpha)\} \\ &= \{T_{10} \geq \log(v_\alpha)\log(\frac{\theta_1(1-\theta_1)}{\theta_0(1-\theta_0)})^{-1}\} \end{aligned}$$

pour $\alpha = 0,05$ et $\mathbb{P}(T_{10} \geq 6) \leq 0,05$

cas de 2 hypothèses multiples

On teste $H_0 : \theta \in \mathbb{H}_0$ contre $H_1 : \theta \in \mathbb{H}_1$
la statistique est le rapport de vraisemblance

$$V(X_1, X_2....X_n) = \frac{Sup_{\theta\in\mathbb{H}1}L(X_1, X_2....X_n, \theta)}{sup_{\theta\in\mathbb{H}0}L(X_1, X_2....X_n, \theta)}$$

Remarque 1.9. *Quand on hésite dans le choix d'une statistique, l'estimateur $\hat{\theta}$ du maximum de vraisemblance est un bon condidat $sup_{\theta\in\mathbb{H}0}L(X_1, X_2....X_n, \theta) = L(X_1, ...X_n, \hat{\theta})$.*

1.2 Vecteur gaussiens

Définition 1.10. *soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$. X est un vecteur gaussien si et seulement si toute combinaison linéaire de ses coordonnées est une variable aléatoire réelle gaussienne de variance éventuellement nulle.*

Définition 1.11. *(Matrice de covariance) Soit X une variable aléatoire à valeurs dans $\mathbb{R}^d$. on note $X = {}^t(X_1, ..., X_d)$. Les variables aléatoires réelles $X_1, ..., X_d$ sont appelées les composantes de X et leurs lois sont les lois marginales de la loi de X.On définit son espérance si les$X_1, ..., X_d$ sont*

intégrables.

$$\mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ . \\ . \\ \mathbb{E}[X_d] \end{pmatrix}$$

et aussi si $X_1, ..., X_d$ sont de carré intégrable sa variance

$$\sum = Var(X) = \mathbb{E}((X - \mathbb{E}(X)^t(X - \mathbb{E}(X)) = (Cov(X_i, X_j))_{1 \leq i,j \leq d}$$

$$\sum = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_d) \\ Cov(X_2, X_1) & Var(X_2) & \dots & Cov(X_2, X_d) \\ \vdots & \vdots & \ddots & \dots \\ Cov(X_d, X_1) & Cov(X_d, X_2) & \dots & Var(X_d) \end{pmatrix}$$

Proposition 1.12. *Soit X un vecteur gaussien dans $\mathbb{R}^d$. on note m son espérance $m = \mathbb{E}[X] = (\mathbb{E}[X_1], ..., \mathbb{E}[X_n])$ et $\sum$ sa matrice de covariance : pour tout i et j $\sum_{i,j} = Cov(X_i, X_j)$ alors la fonction caractéristique de X est pour tout $u \in \mathbb{R}^d$,*

$$\phi_X(u) = \mathbb{E}[e^{i<u,X>}] = e^{i<u,m>} e^{-^t u \sum u/2}$$

On note $X \sim \mathcal{N}(m, \sum)$.

On a $< u, m >= \sum_{k=1}^{d} u_k m_k$ et ${}^t u \sum u =< u, \sum u >$. Remarques :

1. $\sum = o$,$\phi_X(u) = e^{i<u,m>}$ pour tout $u \in \mathbb{R}^d$,X est constant égale à m
2. les coordonées X_i admettent des moments de tout ordre .
3. comme $\phi_X(u)$ caractérise la loi si X est un vecteur aléatoire de fonction caractéristique $u \in \mathbb{R}^d \longmapsto \mathbb{E}[e^{i<u,X>}] = e^{i<u,m>} e^{-^t u \sum u/2}$ alors X est un vecteur gaussien de moyen m et de matrice de covariance $\sum$

Proposition 1.13. *Soit $X \sim \mathcal{N}(m, \sum)$ un vecteur gaussien dans $\mathbb{R}^d$. Soit $A \in \mathcal{M}_{kd}(\mathbb{R})$ .Soit $b \in \mathbb{R}^k$ alors*

$$AX + b \sim \mathcal{N}(Am + b, A \sum {}^t A)$$

Théorème 1.14. *$X \sim \mathcal{N}(m, \sum)$. Les coordonnées de X sont indépendantes, si est seulement si $\sum$ est diagonale.*

Remarque 1.15. *c'est faux si X n'est pas un vecteur gaussien ! Par extension, si (X, Y) est un vecteur gaussien avec X de dimension d et Y de dimension k, X et Y sont indépendantes si est seulement si $Cov(X_i Y_j) = 0$ pour tout $i \leqslant d$ et $j \leqslant k$.*

Théorème 1.16. *(TCL multidimensionnel).Soit $(X_n)_{n \in \mathbb{N}}$ une suite de vecteurs aléatoires iid à valeurs dans $\mathbb{R}^d$, de moyenne m et de matrice de covariance $\sum$ alors*

$$\sqrt{n}(\overline{X}_n - m) \Longrightarrow \mathcal{N}(0, \sum)$$

.

1.2.1 Normes de vecteurs gaussiens centrés

1.2.2 Théorème de Cochran

Théorème 1.17. *soit $X \sim \mathcal{N}(0, I_n)$ où I_n est la matrice identité de taille n, ainsi que $E_1, E_2, ...E_k$ espaces des sous espaces vectoriels deux à deux orthogonaux de $\mathbb{R}^n$ .On note π_{Ej} la projection orthogonale sur E_j où $j \in \{1, 2, ...k\}$.*
La famille de vecteurs gaussiens $(\pi_{Ej}X)_{j\leq k}$ est une famille indépendante et pour chaque j

$$\| \pi_{Ej}X \|_2^2 \sim \chi^2_{dim(E_j)}$$

Chapitre 2

Le test du chi-deux

2.1 Principe du test

Sous P_θ on dispose d'une observation de l'échantillon $X_1(\omega) = x_1, X_2(\omega) = x_2.....X_n(\omega) = x_n$ et on doit prendre une décision concernant la vrai loi des X_i. Dans le test de Khi-2 on teste :

H_0 :la loi des X_i est η
contre

H_1 :la loi des X_i est une autre loi .

Après il faut :

1. Calculer algébriquement la distance entre les données observées et les données théoriques.
2. Déterminer le nombre de degrés de liberté du problème à partir du nombre de classes, et à l'aide d'une table de χ^2, déduire en tenant compte du nombre de degrés de liberté la distance critique qui a une probabilité de dépassement égale à le risque

Donc le but de ce test ,c'est de comparer une distribution théorique d'un caractère à une distribution observée. Pour cela, le caractère doit prendre un nombre fini de valeurs, ou bien ces valeurs doivent être rangées en un nombre fini de classes.

Exemple 2.1. *On veut tester si un dé est truqué, on le lance $n = 100$ fois, on recueille les résultats $X_1, X_2.....X_{100}$ qui sont indépendantes et identiquemet distribuées (iid) de loi $\lambda = Unif(\{1, 2, 3, 4, 5, 6\})$ pour j de 1 à 6 $p_j = P(X_j = j)$ et*

$$N_j = \sum_{i=1}^{n} \mathbb{1}_{X_i=j}$$

le nombre de fois où on a tiré j donc $N_j \sim \mathcal{B}(n, p_j)$ alors on peut prendre $\sum_{i=1}^{n} (N_j - \frac{n}{6})^2 \frac{6}{n}$ comme statistique de test

2.2 Test du rapport de vraisemblance sur les paramétres d'une multinomiale

Soit les $X_1, X_2....X_n$ iid de loi inconnue porté par $a_1, a_2, ...a_k$ et $p = (p_1, p_2, ...p_k)$ avec $p_j = P(X_j = a_j)$ pour chaque j. On veut savoir si la vraie répartition $p = (p_1, p_2, ...p_k)$ est égale à une répartition $q = (q_1, q_2, ...q_k)$ fixée.

On teste $H_0 : p = q$ contre $H_1 : p \neq q$,On calcule la vraisemblance $L(X_1, X_2....X_n, p_1, p_2, ...p_k) = \prod_{j=1}^{k} p_j^{N_j}$ avec $N_j = \sum_{i=1}^{n} \mathbb{1}_{X_i=a_j}$ le nombre de fois où on a observé a_j. Après on cherche à atteindre le maximum de vraisemblance puisque la fonction logarithme est strictement croissante, la vraisemblance et la log-vraisemblance atteignent leur maximum au même point

$$\log(L(X_1, X_2....X_n, p_1, p_2, ...p_k)) = \log(\prod_{j=1}^{k} p_j^{N_j})$$

donc

$$\log(L(X_1, X_2....X_n, p_1, p_2, ...p_k)) = \sum_{j=1}^{k} N_j * \log(p_j)$$

De plus la recherche du maximum de vraisemblance nécessite généralement de calculer la dérivée de la log-vraisemblance pour cela on va utiliser la méthode des multiplicateurs de Lagrange qui permet de trouver les points stationnaires (maximum, minimum...) d'une fonction dérivable d'une ou plusieurs variables, sous contraintes.
On maximise log-vraisemblace sous contrainte $p_1 + p_2..... + p_k = 1$

$$\mathcal{L}(p_1, p_2, ...p_k, \lambda) = \sum_{j=1}^{k} N_j * \log(p_j) + \lambda(\sum_{j=1}^{k} p_j - 1)$$

on calcule l'équation :

$$\nabla\mathcal{L}(\hat{p}, \lambda) = 0 \Leftrightarrow \frac{\partial\mathcal{L}}{\partial p_j} = 0 \text{ et } \frac{\partial\mathcal{L}}{\partial\lambda} = 0 \Leftrightarrow \forall j \in \{1, 2....k\}\ p_j = \frac{-N_j}{\lambda}\ et \sum_{j=1}^{k} p_j = 1$$

Comme $\sum_{j=1}^{k} p_j = 1$, on a $-\frac{\sum_{j=1}^{k} N_j}{\lambda} = 1$ donc $-\frac{n}{\lambda} = 1$ et donc $\lambda = -n$
On calcule le rapport de vraisemblance $\hat{p} = (\frac{N_1}{n}, \frac{N_2}{n}, ...\frac{N_k}{n})$ d'où $p_j = \frac{N_j}{n}$.

$$V = \frac{L(X_1, X_2....X_n, \hat{p})}{L(X_1, X_2....X_n, q)}$$

$$V = \frac{\prod_{j=1}^{k} \frac{N_j}{n^{N_j}}^{N_j}}{\prod_{j=1}^{k} q_j^{N_j}}$$

$$\log(V) = \log(\prod_{j=1}^{k}(\frac{N_j}{nq_j})^{N_j}) = \sum_{j=1}^{k} N_j \log(\frac{N_j}{nq_j})$$

sous H_0 N_j est proche de nq_j donc $N_j \sim \mathcal{B}(n, q_j)$, par un DL on peut écrire

$$\log(\frac{N_j}{nq_j}) \approx \frac{N_j}{nq_j} - 1$$

donc

$$\log(V) \approx \sum_{j=1}^{k} N_j \frac{N_j - nq_j}{nq_j} = \sum_{j=1}^{k} \frac{(N_j - nq_j)^2}{nq_j}$$

car

$$\sum_{j=1}^{k} -nq_j \frac{N_j - nq_j}{nq_j} = -(\sum_{j=1}^{k} N_j - n\sum_{j=1}^{k} q_j) = n - n = 0$$

on prend comme statistique la différence entre effectifs observés et effectif théorique pondéré par les effectifs théoriques

$$T_n = \sum_{j=1}^{k} \frac{(N_j - nq_j)^2}{nq_j}$$

Théorème 2.2. *Si* $(N_1, N_2, ...N_k) \sim Multinomiale(n, p_1, p_2, ...p_k)$*pour chaque* $n \in \mathbb{N}^*$ *alors sous* H_0 *:*

$$T_n = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j} \Longrightarrow \chi^2(k-1)$$

converge en loi

Démonstration. Posons $N = (N_1, ..., N_k)$ où $N_j = \sum_{i=1}^{n} \mathbb{1}_{X_i=j}$ alors on applique le théorème centrale limite multidimensionnel à $N = \sum_{i=1}^{n} X^{(i)}$ avec $X^{(i)} = (\mathbb{1}_{X_i=1}, ...\mathbb{1}_{X_i=k})$

$$Z_n = \sqrt{n}(\frac{N}{n} - p) \longrightarrow \mathbb{N}(0, \sum) \text{ où}$$

$$\sum = \begin{pmatrix} Var(\mathbb{1}_{X_1=1}) & Cov(\mathbb{1}_{X_1=1}, \mathbb{1}_{X_1=2}) & \dots & Cov(\mathbb{1}_{X_1=1}, \mathbb{1}_{X_1=k}) \\ Cov(\mathbb{1}_{X_2=1}, \mathbb{1}_{X_1=1}) & Var(\mathbb{1}_{X_2=2}) & \dots & Cov(\mathbb{1}_{X_2=2}, \mathbb{1}_{X_k=2}) \\ \vdots & \vdots & \ddots & \dots \\ Cov(\mathbb{1}_{X_k=k}, \mathbb{1}_{X_1=k}) & Cov(\mathbb{1}_{X_k=k}, \mathbb{1}_{X_2=k}) & \dots & Var(\mathbb{1}_{X_k=k}) \end{pmatrix}$$

on sait que $X_i \sim Ber(p_i)$. Donc $Var(\mathbb{1}_{X_1=1}) = p_1(1-p_1)$ et

$$Cov(\mathbb{1}_{X_1=1}, \mathbb{1}_{X_1=2}) = \mathbb{E}(\mathbb{1}_{X_1=1}\mathbb{1}_{X_1=2}) - \mathbb{E}(\mathbb{1}_{X_1=1})\mathbb{E}(\mathbb{1}_{X_1=2}) = -p_1p_2$$

$\mathbb{E}(\mathbb{1}_{X_1=1}\mathbb{1}_{X_1=2}) = 0$ car $\mathbb{1}_{X_1=1}\mathbb{1}_{X_1=2} = \mathbb{1}_{X_1=1 \cap X_1=2} = 0$

Alors

$$\sum = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_k \\ -p_2p_1 & p_2(1-p_2) & \dots & -p_2p_k \\ \vdots & \vdots & \ddots & \dots \\ -p_kp_1 & -p_kp_2 & \dots & p_k(1-p_k) \end{pmatrix}$$

Posons $\triangle = \begin{pmatrix} \frac{1}{\sqrt{p_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{p_2}} & \dots & 0. \\ \vdots & \vdots & \ddots & \dots \\ 0 & 0 & \dots & \frac{1}{\sqrt{p_k}} \end{pmatrix}$ une matrice diagonale alors

$$\triangle Z_n \Longrightarrow \mathbb{N}(0, \triangle \sum {}^t\triangle) \quad (proposition 1.13)$$

$$\triangle \sum {}^t\triangle = \begin{pmatrix} \frac{1}{\sqrt{p_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{p_2}} & \dots & 0. \\ \vdots & \vdots & \ddots & \dots \\ 0 & 0 & \dots & \frac{1}{\sqrt{p_k}} \end{pmatrix} \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_k \\ -p_2p_1 & p_2(1-p_2) & \dots & -p_2p_k \\ \vdots & \vdots & \ddots & \dots \\ -p_kp_1 & -p_kp_2 & \dots & p_k(1-p_k) \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{p_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{p_2}} & \dots & 0. \\ \vdots & \vdots & \ddots & \dots \\ 0 & 0 & \dots & \frac{1}{\sqrt{p_k}} \end{pmatrix}$$

donc

$$\triangle \sum {}^t\triangle = \begin{pmatrix} \sqrt{p_1}(1-p_1) & -\sqrt{p_1}p_2 & \dots & -\sqrt{p_1}p_k \\ -\sqrt{p_2}p_1 & \sqrt{p_2}(1-p_2) & \dots & -\sqrt{p_2}p_k \\ \vdots & \vdots & \ddots & \dots \\ -\sqrt{p_k}p_1 & -\sqrt{p_k}p_2 & \dots & \sqrt{p_k}(1-p_k) \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{p_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{p_2}} & \dots & 0. \\ \vdots & \vdots & \ddots & \dots \\ 0 & 0 & \dots & \frac{1}{\sqrt{p_k}} \end{pmatrix}$$

$$\triangle \sum {}^t\triangle = \begin{pmatrix} 1-p_1 & -\sqrt{p_1p_2} & \dots & -\sqrt{p_1p_k} \\ -\sqrt{p_2p_1} & 1-p_2 & \dots & -\sqrt{p_2p_k} \\ \vdots & \vdots & \ddots & \dots \\ -\sqrt{p_kp_1} & -\sqrt{p_kp_2} & \dots & 1-p_k \end{pmatrix} = \sum{}'$$

$\sum{}'$ est une matrice de projection de plus $\sum{}'$ est symétrique donc c'est une projection orthogonale.

On peut alors montrer que le noyau est de dimension 1 et $\sum{}' = {}^t\mathbb{O} \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & 0 \end{pmatrix} \mathbb{O}$ avec $\mathbb{O}$ matrice orthogonale donc on projette sur un espace de dimension $k-1$, le thm de Cochran permet alors de conclure que $\| \triangle Z_n \|^2 \Longrightarrow \chi^2(k-1)$ ∎

Lemma 2.3. *soit $\alpha \in]1, 0[$ le test de région de rejet est $T_n > q_{\chi^2(k-1)(1-\alpha)}$ est un test de niveau*

asymptotique $1-\alpha$

Démonstration. sous H_1 $T_n \longrightarrow \infty$ donc T_n prend des valeurs plus grandes sous H_1 que sous H_0 donc la zone de rejet s'écrit sous la forme suivante $R = \{T_n > C_\alpha\}$ et on cherche C_α au niveau $1-\alpha$ sous H_0 telle que

$$\mathbb{P}(R) = \alpha$$

donc

$$1 - \mathbb{P}(T_n \leq C_\alpha) = \alpha$$

donc

$$\mathbb{P}(T_n \leq C_\alpha) = 1 - \alpha$$

donc

$$C_\alpha = q_{\chi^2(k-1)(1-\alpha)}$$

∎

Remarque 2.4. *l'approximation de la loi T_n sous H_0 par une loi de χ^2 est bonne lorsque les effectifs verifient $np_j \geq 5$ pour tout $k \in [1,k]$*

Exemple 2.5. *on note π la vrai loi du dé et ψ=Unif(1,2,3,4,5,6) la loi d'un dé équilibré on teste $H_0 : \pi = \psi$ contre $H_0 : \pi \neq \psi$ Par défaut on accepte l'idée que le dé est équilibré car s'il n'est pas équilibré, on ne connaitrait pas la loi sous H_0 de la statistique . On lance 100 fois le dé, et on note N_i le nombre de fois où il donne le résultat i. La statistique du test est $T_n = \sum_{j=1}^{6} \frac{(N_j - \frac{n}{6})^2}{\frac{n}{6}}$.*
On utilise le théorème de convergence des multinomiales vers les χ^2, sous H_0 la loi de T_n est proche de $\chi^2(5)$ car $np_j(1-p_j) = 100\frac{1}{6}(1-\frac{1}{6}) = 13.89 \geq 5$ pour tout $j \in [1,6]$

SousH_1 les effectifs observés N_i s'éloignent des effectifs théoriques $\frac{1}{6}$ donc T_n augmente, la table de la loi χ^2 donne la région de rejet $R = \{T_n > 11,07\}$ au niveau $1-\alpha$ avec $\alpha = 0,05$

sur les 100 lancers on observe

$$N_1(\omega) = 16, \quad N_2(\omega) = 20, \quad N_3(\omega) = 19, \quad N_4(\omega) = 10, \quad N_5(\omega) = 17, \quad N_6(\omega) = 18,$$

La valeur observée de la statistique est donc

$$T_{100}(\omega) = \frac{(16-\frac{100}{6})^2 + (20-\frac{100}{6})^2 + (19-\frac{100}{6})^2 + (10-\frac{100}{6})^2 + (17-\frac{100}{6})^2 + (18-\frac{100}{6})^2}{\frac{100}{6}} = 3,8$$

C'est au dessous de 11.07 *alors on conserve H_0.*
La p-valeur de ce test est très élevée $\mathbb{P}_{H0}(T_n \geqslant 3.8) \approx 0.5$ et elle signifie que ce dé est très équilibré.

Chapitre 3

Application

On souhaite savoir si les entrées à l'hôpital pour une certaine maladie sont réparties au hasard dans l'année ou bien si certains mois sont plus propices à la maladie. On examine le mois d'entrée d'un échantillon de 120 porteurs de la maladie étudiée. Les résultats sont :

Mois d'entrée	1	2	3	4	5	6	7	8	9	10	11	12
Nombre d'entrées	18	16	8	10	6	4	4	9	11	10	12	12

Peut-on affirmer, au risque 0.01 que "les entrées ne se font pas au hasard dans l'année" ?

Soit X la var égale au le mois d'entrée à l'hôpital. Par l'énoncé, on observe la valeur de X sur chacun des n individus (porteurs de la maladie) d'un échantillon avec $n = 120$: $(x_1, ..., x_n)$. Ces valeurs sont regroupées en $k = 12$ classes : $C_1 = 1, \quad C_2 = 2, ..., \quad C_{12} = 12$, avec pour effectifs respectifs : $n_1 = 18, \quad n_2 = 16, ..., \quad n_{12} = 12$. Dire que les entrées se font au hasard dans l'année signifie que X suit la loi uniforme $U(\{1, ..., 12\})$: $\forall i \in \{1, ...n\} \ \mathbb{P}(X = i) = \dfrac{1}{12}$
on teste H_0 : "X suit la loi uniforme $U(\{1, ..., 12\})$"
contre
H_1 : "X ne suit pas la loi uniforme $U(\{1, ..., 12\})$".
Soit R une var suivant la loi uniforme $U(\{1, ..., 12\})$. On a $p_i = \mathbb{P}(R \in C_i) = \dfrac{1}{12}$ pour tout $i \in \{1, ..., 12\}$.

```
nb = c(18, 16, 8, 10, 6, 4, 4, 9, 11, 10, 12, 12)
proba = rep(1 / 12, 12)
chisq.test(nb, p = proba)$p.value
```

Cela renvoie : 0.04267211
Comme p-valeur $\in]0.01, 0.05]$, le rejet de H_0 est significatif .

Annexe

Définition 3.1. *(la loi de Bernoulli)Une variable aléatoire X suivant la loi de Bernoulli de probabilité p si $\mathbb{P}(X = k) = p^k(1-p)^{1-k}$ pour tout $k \in \{0, 1\}$*
C'est à dire

$$\mathbb{P}(X = x) = \begin{cases} p & \text{si } k = 1, \\ 1-p & \text{si } k = 0, \\ 0 & \text{sinon.} \end{cases}$$

Proposition 3.2.

$$\mathbb{E}(X) = p \quad Var(x) = p(1-p)$$

Définition 3.3. *(Loi binomiale) La loi binomiale, de paramètres n et p, est la loi de probabilité d'une variable aléatoire X telle que : $X = Y_1 + Y_2 + ... + Y_n$, où $Y_1, Y_2, , , Y_n$,sont des variables aléatoires indépendantes de loi de Bernoulli de même paramètre p et sa fonction de masse est donnée par*

$$\mathbb{P}(X = k) = C_n^k p^k (1-p)^{n-k}$$

Proposition 3.4.

$$\mathbb{E}(X) = np \quad Var(x) = np(1-p)$$

Définition 3.5. *Soient $k \in \mathbb{N}$ et $X_1, ..., X_k$ des variables aléatoires indépendantes de loi normale centrée réduite $\mathcal{N}(0, 1)$. On appelle loi du χ^2 à k degrés de libertés, et on note $\chi^2(k)$, la loi de $X_1^2 + ... + X_k^2$.*

Proposition 3.6. *Si $X \sim \chi^2(k)$, alors $\mathbb{E}(X) = k$ et $Var(X) = 2k$.*

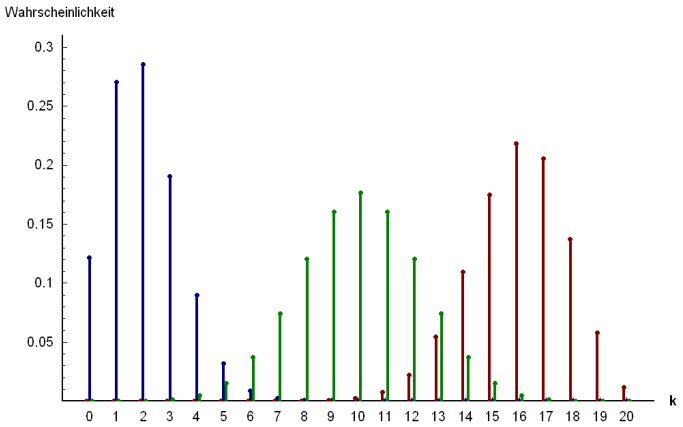

FIGURE 3.1 – Diagrammes en bâtons de trois fonctions de masse de lois binomiales. Les paramètres sont n = 20 et p = 0,1 (en bleu), p = 0,5 (en vert) et p = 0,8 (en rouge).

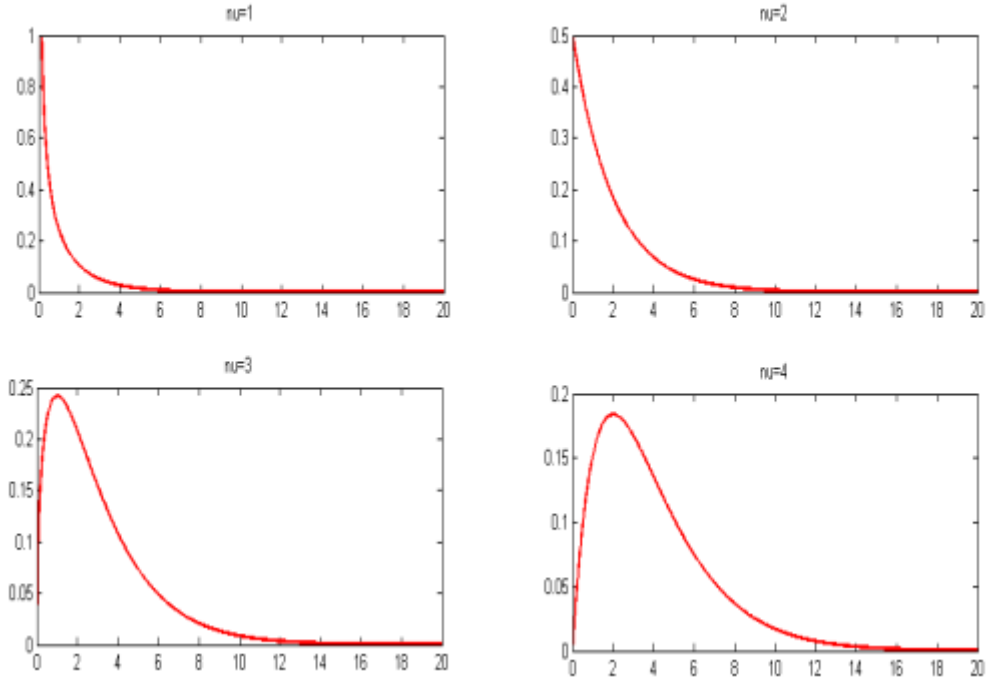

FIGURE 3.2 – Densité de la loi $\chi^2(k)$

```
x <- seq(from=0, to=20, by=2)
y1 <- dchisq(x,1)
y2 <- dchisq(x,2)
y3 <- dchisq(x,3)
y4 <- dchisq(x,4)
plot(x,y1,"l",col="red",xlim=c(0,20),ylim=c(0,0.05),ylab="")
plot(x,y2,"l",col="red",xlim=c(0,20),ylim=c(0,0.05),ylab="")
plot(x,y3,"l",col="red",xlim=c(0,20),ylim=c(0,0.05),ylab="")
plot(x,y4,"l",col="red",xlim=c(0,20),ylim=c(0,0.05),ylab="")
```

Conclusion

Les statistiques, dans le sens populaire du terme, traitent des populations. En statistique descriptive, on se contente de décrire un échantillon à partir de grandeurs comme la moyenne, la médiane, l'écart type, la proportion, la corrélation, etc. C'est souvent la technique qui est utilisée dans les recensements La statistique mathématique repose sur la théorie des probabilités. Des notions comme la mesurabilité ou la convergence en loi y sont souvent utilisées. Mais il faut distinguer la statistique en tant que discipline et la statistique en tant que fonction des données.

Bibliographie

[1] https ://e-formation.uha.fr/pluginfile.php/16418/mod$_r esource/content/0/15Chi2.pdf$

[2] https ://www.math.univ-paris13.fr/ tournier/fichiers/agreg/statistiques.pdf

[3] https ://www.lpsm.paris/pageperso/akakpo/documents/PolyTests2017.pdf

[4] https ://fr.wikipedia.org/wiki/LoideBernoulli

[5] https ://fr.wikipedia.org/wiki/Loibinomiale

[6] https ://chesneau.users.lmno.cnrs.fr/adequation-R.pdf