L2 Mention Informatique

UE Probabilités

Chapitre 5 : Bases de statistiques

Notes de cours rédigées

par

Régine André-Obrecht et Julien Pinquier

Université Paul Sabatier

TOULOUSE III

I. Introduction

Quelques problèmes typiques :

- détermination de la loi de répartition d'une variable aléatoire (ou d'un système de variables aléatoires) d'après des données statistiques,
- vérification de la vraisemblance des hypothèses,
- recherche des paramètres d'une loi de répartition.

II. Fonction de répartition statistique (empirique)

Soit une variable aléatoire X, dont la loi de répartition est inconnue. On peut découvrir cette loi à partir de l'expérience aléatoire ou vérifier expérimentalement l'hypothèse que X suit telle ou telle loi.

Une série d'expériences indépendantes permet de recueillir des observations, à partir desquelles on obtient des réalisations de la variable X, $x_i = X(w_i)$.
L'ensemble des valeurs observées $(x_1, ..., x_n)$ est appelé « **ensemble statistique simple** » ou « **suite statistique simple** ».

<u>Remarque</u> : Parfois on considère une suite de variables indépendantes $X_1, ..., X_n$ de même loi que X. Dans ce cas X est appelée une **variable aléatoire parente** et $(X_1, ..., X_n)$ l'**échantillon** de taille **n** issu de X.

Le terme « échantillon » désigne un ensemble important d'éléments (d'individus) extraits d'une population (collectivité statistique). Il s'agit d'une variable aléatoire dont la valeur change d'un individu à l'autre. Cependant pour se faire une idée de la répartition de cette variable aléatoire il n'est pas indispensable d'étudier chaque individu de la collectivité ; on peut étudier un certain échantillon suffisamment grand : on parle alors d'**inférence statistique** (exemple : sondage). L'ensemble dans lequel on prélève l'échantillon s'appelle en statistique le **grand ensemble**. On suppose que le nombre d'éléments (d'individus) **N** du grand ensemble est très important, et le nombre d'éléments **n** d'un échantillon est limité.

Lorsque N est suffisamment grand les propriétés des distributions échantillonnées (statistiques) et des caractéristiques ne dépendent pratiquement pas de N ; d'où l'idéalisation mathématique que l'ensemble dont sont issus les échantillons est de taille infinie. Alors la loi de X correspond à la loi déterminée par le « grand ensemble ».

Définition On appelle **fonction de répartition statistique** (ou empirique) de la variable aléatoire X la fréquence de l'événement X < x dans les données statistiques :

$$F^*(x) = P(X < x) = \frac{nb_{exp}(X < x)}{n}.$$

- $F^*(x)$ est croissante,
- $F^*(x) = 0$ si $x \le \min(x_i)$,
- $F^*(x) = 1$ si $x > \max(x_i)$.

Si chaque valeur d'une variable aléatoire a été observée exactement 1 fois, le saut (la marche) de la fonction de répartition statistique (pour chaque valeur observée) est égal(e) à 1/n.

Lorsque n augmente la fonction de répartition $F^*(x)$ tend (en probabilité) vers la vraie fonction de répartition F(x) de la variable aléatoire X.

III. Histogramme

Si le nombre d'observations est important, les données statistiques peuvent être représentée sous une forme plus compacte et ordonnée. Pour cela, on divise la gamme des valeurs observées de X en **intervalles** (ou rangs), et on compte le nombre de valeurs correspondant à chaque rang.

En divisant ce nombre n_i^* par le nombre total d'observations n, on trouve la **fréquence** correspondant à un rang donné : $p_i^* = \frac{n_i}{n}$.

On a $\sum_{i=1}^{k} p_i^* = 1$ avec k le nombre de rangs.

Définition On appelle **suite statistique** le tableau suivant…

Tableau 1: Suite statistique.

I_i	$[r_1, r_2]$	$[r_2, r_3]$	…	$[r_k, r_{k+1}]$
p_i^*	p_1^*	p_2^*	…	p_k^*

avec : I_i le $i^{ème}$ rang, $[r_i, r_{i+1}]$ ses limites et p_i^* sa fréquence.

Un **histogramme** est une forme de représentation graphique de la suite statistique. Sur l'axe des abscisses se trouvent les intervalles de classement. Pour chaque intervalle, la hauteur du rectangle est égale à la division de la fréquence du rang par sa longueur.

IV. Moments statistiques

Définition : La **moyenne statistique** (empirique) d'une variable aléatoire est définie par :

$$m^*(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

avec x_i la valeur observée dans la $i^{ème}$ expérience et n le nombre d'expériences.

Définition : Sachant que $\mathrm{var}(X) = E\left((X - E(X))^2\right)$, la **variance statistique** (empirique) est égale à :

$$s^*(X) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - m^*(X)\right)^2$$

De la même façon, on définit les moments statistiques centrés d'ordre quelconque.

V. Critères de conformité

Il s'agit de mesurer la **conformité des répartitions théorique et statistique**.

Supposons que la fonction de répartition statistique soit approchée par la courbe théorique. Même si la courbe théorique est bien choisie, certains écarts entre celle-ci et la fonction statistique sont inévitables.

<u>Question</u> : les écarts sont-ils dus au « hasard » (vu le nombre limité d'observations), ou sont-ils essentiels ? (En d'autres mots avons-nous alors mal choisi la loi théorique !).

Prenons comme hypothèse **H** : « *la variable X a $F_X(x)$ pour fonction de répartition* ».

Pour adopter ou rejeter l'hypothèse H, on considère une certaine grandeur **U** caractérisant la différence entre les fonctions de répartitions statistique et théorique.

<u>Exemple</u> : $U = \max |F^*(x) - F(x)|$

Supposons que nous connaissons la loi de répartition. Dans ce cas U est une variable aléatoire, du fait que F* est une fonction de X et on peut calculer la probabilité de $U \geq u$.

<u>Remarque</u> : Si « sachant F », on a trouvé que :

- $P(U \geq u)$ est trop faible → c'est une raison de rejeter l'hypothèse,
- $P(U \geq u)$ est assez grande → on dit que H n'est pas en contradiction avec les données expérimentales,
- $P(U \geq u)$ est trop proche de 1 → peut être on a truqué les données...

Un test de conformité : méthode du χ^2

Supposons que l'on ait une suite statistique (cf. Tableau 1).
Soit p_i la probabilité théorique $P(X \in I_i)$.

La valeur du test de χ^2 est donnée par :

$$\chi^2 = \sum_{i=1}^{k}\frac{n}{p_i}(p_i^* - p_i)^2 = n\sum_{i=1}^{k}\frac{(p_i^* - p_i)^2}{p_i} = \sum_{i=1}^{k}\frac{(n_i^* - np_i)^2}{np_i}.$$

Il s'agit de la somme des carrés de la différence entre les fréquences théorique et statistique pondérées. « Le poids est plus grand si la probabilité théorique est plus faible ».

Il existe des tables de loi du χ^2 (cf. **Erreur ! Source du renvoi introuvable.**). La loi dépend du nombre de degrés de liberté **r** du χ^2.
Ce degré est évalué dans le test statistique par $r = k - s$.
avec :

- **k**, le nombre d'intervalles,
- **s**, le nombre de contraintes (équations) imposées.

<u>Exemples de contraintes :</u>

1- La contrainte $\sum_{i=1}^{k} p_i^* = 1$ est toujours imposée. Si c'est la seule, s = 1.

2- Il est souvent exigé que la moyenne (espérance) théorique coïncide avec la moyenne statistique : $E(X) = m^*(X)$, et dans ce cas, s = 2.

3- Si la variance théorique doit coïncider avec la variance statistique, alors s = 3.

Et ainsi de suite pour les moments d'ordre supérieur…

<u>Méthode de recherche dans la table</u> : On cherche dans la ligne qui correspond à r, à la valeur la plus proche de $u = \chi^2$ calculée, la valeur de $p = P(\chi^2 \geq u)$.

Unité de cours Probabilités - Exercices –
Chapitre 5 : Bases statistiques

Exercice 1*

Prendre les 20 premières lignes et les 4 premières colonnes de la table de nombres au hasard. Considérer les 4 chiffres de chaque ligne comme les 4 décimales d'un nombre $X \in [0,1]$.

1- Faire une suite statistique : compléter le tableau suivant :

I_i	[0 ; 0,2]	[0,2 ; 0,4]	[0,4 ; 0,6]	[0,6 ; 0,8]	[0,8 ; 1]
p_i^*	?	?	?	?	?

2- Dessiner l'histogramme.

Comparaison avec la loi uniforme sur [0 ; 1] en utilisant le critère du χ^2.

3- Calculer la valeur du χ^2.
4- Combien de degrés de liberté doit-on prendre ?
5- Quelle est (approximativement) la probabilité d'une telle valeur de χ^2 selon la table du χ^2 ?

Exercice 2*

Pour un échantillon de taille m=100, on compte 2 valeurs dans l'intervalle [-2, -1[, 8 valeurs dans l'intervalle [-1, 0[, 20 valeurs dans l'intervalle [0, 1[, 35 valeurs dans l'intervalle [1, 2[, 30 valeurs dans l'intervalle [2, 3[, et 5 valeurs dans l'intervalle [3, 4[.

1- Donner la suite statistique associée et tracer l'histogramme.
2- Estimer la plus grande et plus petite moyenne possible de l'échantillon.
3- Sous les deux hypothèses H1={la loi de la variable aléatoire est une loi normale N(1,1)} et H2={la loi de la variable aléatoire est une loi normale N(2,1)}, calculer pour chaque hypothèse les probabilités théoriques de chaque intervalle, en utilisant les tables de la loi centrée réduite N(0,1) (chapitre 3).
4- Calculer le critère du χ^2 pour chaque hypothèse et conclure.

Exercice 3

Prendre les 4 dernières colonnes (37-40) et les lignes 16-25 de la table de nombres au hasard. Considérer les 4 chiffres de chaque ligne comme les 4 décimales d'un nombre appartenant à [0,1] . On peut observer que, bien que la loi soit supposée uniforme, la plupart des valeurs ne dépasse pas 0.5. En particulier 5 valeurs sont concentrées dans [0, 0.2]. Il est possible que la loi sous jacent soit définie par la densité f_2 suivante :

$$f_2(x) = 0 \quad si \quad x \notin [0,1]$$
$$f_2(x) = 2.5 \quad si \quad x \in [0,0.2]$$
$$f_2(x) = 0.625 \quad si \quad x \in [0.2,1]$$

1- Tracer cette fonction de densité
Au cours de cet exercice, nous allons essayer de comparer par le test du χ^2, les deux hypothèses H_1 : loi uniforme sur [0,1] et H_2 loi définie par la densité f_2 sur [0,1].
2- Faire une suite statistique en complétant le tableau suivant pour chacune des hypothèses:

I_i	[0 ; 0,2]	[0,2 ; 0,4]	[0,4 ; 0,6]	[0,6 ; 0,8]	[0,8 ; 1]
p_i^*	?	?	?	?	?

3- Quelles sont les probabilités théoriques pour chacun des intervalles et chacune des hypothèses ?
4- Trouver la valeur du χ^2, noté T_1 (resp. T_2) pour l'hypothèse H_1 (resp. H_2).
5- Quel est le nombre de degrés de liberté ?
6- Pour chacune des hypothèses, trouver approximativement dans la table du χ^2 :
   a. Les valeurs u_{11},p_{11} et u_{12}, p_{12} telles que
   $$P(\chi^2 \succ u_{11}) = p_{11}, P(\chi^2 \succ u_{12}) = p_{12} \quad et \quad u_{11} \geq T_1 \geq u_{12}$$
   b. Les valeurs u_{21},p_{21} et u_{22}, p_{22} telles que
   $$P(\chi^2 \succ u_{21}) = p_{21}, P(\chi^2 \succ u_{22}) = p_{22} \quad et \quad u_{21} \geq T_2 \geq u_{22}$$
7- Conclure sur l'hypothèse la plus vraisemblable.