Lecture Notes: Proba-Stats

By Thibaud Paulin

Taught by Prof. Todorov Konstantin

M1 IASD,

Outline

Pour exam

3 parties:

1: On doit donner la stat à partir d'une loi ou inversement (loi non centrée-réduite)

2: Mise en place d'un test statistique (accepter ou rejeter H_0) Tests

3: QCM

Chapter 1: Lois usuelles

Definition 1.0.1

Loi uniforme discrète : Si la variable aléatoire X suit une loi uniforme discrète, alors chaque événement a autant de chances d'arriver. Si on a r possibilités, alors chaque événement a une probabilité de $\frac{1}{r}$.
$E(X) = 1\frac{1}{r} + 2\frac{1}{r} + ... + r\frac{1}{r} = \frac{r+1}{2}$
$V(X) = \frac{r^2-1}{12}$

Definition 1.0.2

Loi de Bernoulli $\mathcal{B}(p)$:
Soit p un entier qui est soit 0 soit 1. On note $P(X = 1) = p$ et $P(X = 0) = 1 - p$.
De manière générale, $P(X = x) = p^x(1 - p)^{1-x}$.
$E(X) = p$
$\text{Var}(X) = p(1 - p)$

Definition 1.0.3

Le procesuss de Bernoulli et la loi binomiale $\mathcal{B}(n, p)$:
C'est la répétition d'expérience de Bernoulli, chacune des n expériences ont une probabilité p d'arriver.
$P(X = x) = p(x) = \binom{n}{x} p^x(1 - p)^{n-x}$
$E(X) = np$
$V(X) = np(1 - p)$

La loi de poisson ne sera pas à l'exam

Definition 1.0.4

Loi continue uniforme $\mathcal{U}([a, b])$:
$P(X = x) = 0$. En effet, si la probabilité est de $\frac{1}{n}$, avec un n qui tend vers l'infini, on converge vers 0.
Elle a pour densité définie sur $\mathbb{R}$: $f(x) = \begin{cases} \frac{1}{b-a} & \text{si a} \leq x \leq b \\ 0 & \text{sinon} \end{cases}$

Definition 1.0.5

Loi de Gauss ou loi normale $\mathcal{N}(\mu, \sigma^2)$:
Elle a pour densité définie sur $\mathbb{R}$: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$

Pour $\mu = 0$ et $\sigma^2 = 1$, on parle de loi de Gauss centrée réduite $\mathcal{N}(0,1)$

Proposition 1.0.1

Si $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$ alors $Z = \frac{X-\mu}{\sigma} \rightsquigarrow \mathcal{N}(0,1)$. On peut donc toujours se ramener à une loi de Gausse centrée réduite

Imaginons qu'on veuille demander à chacun des habitants de Montpellier s'ils sont favorables à une proposition de loi. Alors la première personne à réaliser l'enquête va demander à $x_1^1, x_2^1, ..., x_n^1$.Un autre enquêteur va demander à $x_1^2, x_2^2, ..., x_n^2$ etc. Chaque enquêteur aura ses propres résultats car ils interrogeront sûrement des personnes différentes.
Ainsi, tous les x_i^1 sont une variable aléatoire $X_1 \rightsquigarrow$ un processus de Bernoulli. C'est la "loi-mère" . Ele sera identique pour toutes les variables aléatoires X_i.

Definition 1.0.6

Soit$X_1, ..., X_n$ un n-échantillon. On appelle **statistique** sur cette échantillon toute variable aléatoire de la forme $T_n = h(X_1, ..., X_n)$.

Definition 1.0.7

Les tests statistiques ont pour objet de d´ecider sur la base d'un échantillon si une caractéristique de la loi mère répond ou non à une certaine sp´ecification que l'on appelle **hypothèse**.

1.1 Cadre paramétrique

Dans ce cours, nous allons voir le **cadre paramétrique** : les hypothèses portent sur un paramètre inconnu θ ou sur une fonction de ce paramètre $h(\theta)$, correspondant à une caractéristique d'intérêt de la loi.

Ce paramètre peut être la moyenne, la variance etc... θ peut être un réel dans $\mathbb{R}$ ou une intervalle $\theta \in [\theta_1, \theta_2]$, avec $[\theta_1, \theta_2] \subseteq \mathbb{R}$

La loi observée appartient à une famille de lois décrite par la famille de densités de probabilités :

$$\{f(x, \theta) |\ \theta \in \Theta\}$$

La forme fonctionnelle de f est connue, θ est inconnue. La fonction de répartition est notée $F(x, \theta)$. L'ensemble Θ est **l'espace paramétrique**.

Definition 1.1.8

Dans l'approche paramétrique, un test statistique consiste à décider d'accepter ou de rejeter une hypothèse spécifiant que : $\theta \in \Theta_0$, avec $\Theta_0 \subseteq \Theta$.
Cette hypothèse de référence est **l'hypothèse nulle**

Definition 1.1.9

Par opposition, l'**hypothèse alternative** H_1 est l'hyopthèse pour laquelle $\theta \in \Theta \setminus \Theta_0$

On veut donc tester $H_0 : \theta \in \Theta_0$ et $H_1 : \theta \in \Theta_1$.
Suivant la nature de Θ_0 vs Θ_1, trois cas :

1. Hypothèse nulle : simple et hypothèse alternative : simple :
   $H_0 : \theta = \Theta_0$ vs $H_1 : \theta = \Theta_1$
2. Hypothèse nulle : simple et hypothèse alternative : multiple :
   $H_0 : \theta = \Theta_0$ vs $H_1 : \theta \neq \Theta_0$
3. Hypothèse nulle : multiple et hypothèse alternative : multiple :
   $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$

1.2 Test d'une hypothèse simple avec alternative simple

Dans le cas d'hypothèses simples, l'espace paramétrique est $\Theta = \{\theta_0, \theta_1\}$, et le test veut décider : $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$.

Definition 1.2.10

Un test pour H_0 est une règle de décision fondée sur la valeur réalisée t sur un échantillon, d'une statistique T, appelée **statistique du test**, à valeurs dans $\mathbb{R}$.

Maintenant, comment décider si on accepte ou rejette H_0 ? On définit la règle d'acceptation suivante :
- $t \in A$ (avec $A \subseteq \mathbb{R}$), $\Rightarrow$ on accepte H_0
- $t \in \overline{A} \Rightarrow$ on rejette H_0

Definition 1.2.11

La partie A est appelée région d'acceptation du test, et la partie $\overline{A}$ la **région de rejet** du test. Lorsque A est un intervalle, il est appelé **intervalle de confiance** lié au test.

Problem 1.2.1

Une telle règle de décision recèle deux types d'erreurs possibles du fait que la vraie valeur du paramètre est inconnue.

1. Risque de première espèce (faux négatif) :

   ### Definition 1.2.12

   On appelle **risque de première espèce** pour un test, la valeur α telle que : $P_{H_0}(T \in A) = \alpha$, c'est-à-dire la probabilité de rejeter H_0 alors qu'elle est vraie

2. Risque de deuxième espèce (faux positif) :

   ### Definition 1.2.13

On appelle **risque de deuxième espèce** pour un test, la valeur β telle que : $P_{H_1}(T \in A) = \beta$, c'est-à-dire la probabilité d'accepter H_0 alors qu'elle est fausse

Dans un test statistique, on privilégie le risque α (par ex. $\alpha = 0.05$) que l'on fixe a priori. La valeur α est aussi appelée **niveau** ou **niveau de signification du test**.

$\Rightarrow$ **la loi de la statistique de test sour H_0 doit être connue**.

1.3 Construction d'un test

Les étapes de construction d'un test sont les suivantes :

1. Déterminer les hypothèses H_0 et H_1
2. Rechercher une statistique pertinente dont on connaît la loi sous H_0
3. Fixer un niveau α
4. Déterminer l'intervalle de confiance associé à ce niveau, en utilisant la loi de la statistique de test
5. Prendre une décision en considérant la réalisation de la statistique sur l'échantillon

Problem 1.3.2

Comment juger de la pertinence de la statistique choisie ?

Il est naturel de la prendre de telle sorte que la probabilité de rejeter H_0 soit nettement plus élevée sous H_1 que sous H_0

Definition 1.3.14

On appelle **puissance d'un test** la probabiliée de rejeter H_0 alors qu'elle est effectivement fausse, soit :

$$P\left(T \in \overline{A} | H_1\right)$$

La **puissance**, qui est la **capacité à détecter qu'une hypothèse nulle est fausse**, est égale à $1 - \beta$.

Definition 1.3.15

On dit qu'un test est sans biais si sa puissance est supérieure ou égale à son niveau α :

$$P(T \in A | H_1) \geq P(T \in A | H_0)$$

Logiquement, une condition naturelle pour qu'une statistique soit éligible pour tester une hypothèse est qu'elle induise un test sans biais.

1.4 Tests d'hypothèses multiples

La définition de la puissance du test va changer :

$$P\left(T \in \overline{A} | H_1\right)$$

n'a plus de sens si H_0 est multiple

Correction TD2 Ex14 :

105 g pour 100 f la valeur théorique est de $0.51 = \frac{g}{205}$

1. Rejet : nombre de f = nombre de g ? On a donc les variables $X_1, ..., X_n$ avec $X_i = 1$: i-ème g, $X_i = 0$: i-ème f

   Ces variables suivent une loi de Bernoulli avec $P(X_i = 1) = p$ et $P(X_i = 0) = 1 - p$

2. Hypothèses :

   $H_0 : p = p_0 = \frac{1}{2}$ vs $H_1 : p > p_0 = \frac{1}{2}$

3. Définition de l'intérêt $\overline{X} = \frac{1}{n} \sum X_i$: taux de naissance de g dans une population de taille n.

   $\overline{X} \sim B(n, p)$. On va approximer cette variable $\overline{X}$ qui sera donc $\overline{\overline{X}}$ sous H_0

   Donc $\overline{\overline{X}} \sim \mathcal{N}\left(p_0, \frac{p_0(1-p_0)}{n}\right)$

4. Définition de l'intervalle de confiance Posons donc $\alpha = 0.05$ le risque de premier ordre. On cherche donc un seuil s au delà duquel on trouve la zone de rejet. C'est dans cette zone que H_0 sera rejeté car les valeurs seront vraiment très éloignées du "$\frac{1}{2}$" théorique.

   $\alpha = P_{H_0}\left(\overline{\overline{X}} > s\right) = 0.05$ Avant de regarder dans la table la valeur de s, on centre et réduit :

   $\alpha = P(\mathcal{U} > u) = 0.05$ avec $\mathcal{U} \sim \mathcal{N}(0, 1)$

   $\mathcal{U} = \frac{\overline{\overline{X}} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

   $u = \frac{s - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ et en lisant la table 2 on a $u = 1.6449$

   $s - p_0 = 1.6449\sqrt{\frac{p_0(1-p_0)}{n}}$

   $s = 1.6449\sqrt{\frac{p_0(1-p_0)}{n}} + p_0$ avec $p_0 = \frac{1}{2}$ et $n = 79080$

   Ainsi $s = 0.50292$. Décision : la valeur observé de l'échantillon à savoir $p_0(e) = 0.51$
   Or $p_0(e) > s$ qui est dans la zone de rejet ! On rejette donc l'hypothèse nulle H_0

Maintenant, si on veut accepter H_0, on définit une double zone de rejet, à la fois trop inférieure et trop supérieure à $\frac{1}{2}$.

2. Hypothèses
   $H_0 : p = p_0 = \frac{105}{205}$ vs $H_1 : p \neq p_0 = \frac{105}{205}$

3. L'intérêt est le même

   $\overline{\overline{X}} \sim \mathcal{N}\left(p_0, \frac{p_0(1-p_0)}{n}\right)$

4. Définition du risque

   $\alpha = P_{H_0}\left(\overline{\overline{X}} \notin [s_1, s_2]\right)$

   $P_{H_0}\left(\overline{\overline{X}} < s_1 \cup \overline{\overline{X}} > s_2\right)$

   $P\left(\overline{\overline{X}} < s_1\right) = \frac{\alpha}{2}$ et $P\left(\overline{\overline{X}} > s_2\right) = \frac{\alpha}{2}$

   $P\left(s_1 < \overline{\overline{X}} < s_2\right) = 0.95$. On centre et on réduit :

$P\left(\frac{s_1-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < \mathcal{U} < \frac{s_2-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right) = 0.95$

$s_1 = -u\sqrt{\frac{p_0(1-p_0)}{n}} + p_0$ et $s_2 = u\sqrt{\frac{p_0(1-p_0)}{n}} + p_0$ et en lisant la table 2 on a $u = 1.9599$

$s_1 = 0.5087$ et $s_2 = 0.5156$

La zone d'acceptation est donc $IC = [0.5087, 0.5156]$

Décision : la valeur observée $p_0(e) = 0.51 \in IC$, donc on accepte H_0.

Exercice 15 : Effet de la taille de l'échantillon

On jette une pièce, quel est le nombre de faces ? Data : 100 fois $\rightarrow$ 55 faces, 1000 fois $\rightarrow$ 550 faces et 10000 fois $\rightarrow$ 5500 faces

Question : proba "face" $= 0.5$? Selon les différentes datas

(Toujours la même loi de Bernoulli sur les mêmes variables aléatoires, même manière de l'approximer) le seul truc qui change, c'et qu'on va faire $n \times p$ car in s'intéresse au NOMBRE de faces et paux au TAUX de faces. Attention à ce piège de faire un taux à la place du nombre.