Concepts fondamentaux de la statistique

4 A E.S.I.E.A.

E. Claeys

ICUBE/IRMA
Université de Strasbourg

E.S.I.E.A, 2019

Programme du cours

Exemples et problématique

L'état décide d'organiser un référendum sur la légalisation du cannabis. À cette occasion, on interroge n français. On note $X_i = 1$ si l'individu répond « oui » et $X_i = 0$ si l'individu répond « non ». Les individus sont choisis au hasard dans la population en âge de voter.

- Les variables X_i sont-elles indépendantes ?
- Les variables X_i sont-elles identiquement distribuées ?

Exemples et problématique

L'état décide d'organiser un référendum sur la légalisation du cannabis. À cette occasion, on interroge n français. On note $X_i = 1$ si l'individu répond « oui » et $X_i = 0$ si l'individu répond « non ». Les individus sont choisis au hasard dans la population en âge de voter.

- Les variables X_i sont-elles indépendantes ? OUI et NON

$\Rightarrow$ Indépendance = Les X_i ne s'influencent pas entre eux.

- Les variables X_i sont-elles identiquement distribuées ? OUI et NON

$\Rightarrow$ Identiquement distribuées = Les X_i suivent la même loi de probabilité

En réalité, les gens s'influencent entre eux et le vote dépend des caractéristiques décrivant le votant. Pour établir un modèle, il faut parfois simplifier la réalitée et faire des suppositions !

Exemples et problématique

L'état décide d'organiser un référendum sur la légalisation du cannabis. À cette occasion, on interroge n français. On note $X_i = 1$ si l'individu répond « oui » et $X_i = 0$ si l'individu répond « non ». Les individus sont choisis au hasard dans la population en âge de voter. On considère les variables X_i comme indépendantes et identiquement distribués (i.i.d.) suivant une loi de Bernoulli de paramètre $\theta \in [0, 1]$ inconnus.

- Qu'est ce qu'une loi de probabilité ?
- Qu'est ce qu'un paramètre ?
- Qu'est ce qu'une loi de Bernoulli ?

Exemples et problématique

L'état décide d'organiser un référendum sur la légalisation du cannabis. À cette occasion, on interroge n français. On note $X_i = 1$ si l'individu répond « oui » et $X_i = 0$ si l'individu répond « non ». Les individus sont choisis au hasard dans la population en âge de voter. On considère les variables X_i comme indépendantes et identiquement distribués (i.i.d.) suivant une loi de Bernoulli de paramètre $\theta \in [0, 1]$ inconnus.

- Qu'est ce qu'une loi de probabilité ?

$\Rightarrow$ Une loi de probabilité décrit le comportement d'une variable aléatoire à travers un modèle mathématique.

Exemples et problématique

L'état décide d'organiser un référendum sur la légalisation du cannabis. À cette occasion, on interroge n français. On note $X_i = 1$ si l'individu répond « oui » et $X_i = 0$ si l'individu répond « non ». Les individus sont choisis au hasard dans la population en âge de voter. On considère les variables X_i comme indépendantes et identiquement distribués (i.i.d.) suivant une loi de Bernoulli de paramètre $\theta \in [0, 1]$ inconnus.

- Qu'est ce qu'un paramètre ?

$\Rightarrow$ Un paramètre est un élément intervenant dans le modèle mathématique. En informatique cela peut se voir comme un argument nécessaire à une fonction.

Exemples et problématique

L'état décide d'organiser un référendum sur la légalisation du cannabis. À cette occasion, on interroge n français. On note $X_i = 1$ si l'individu répond « oui » et $X_i = 0$ si l'individu répond « non ». Les individus sont choisis au hasard dans la population en âge de voter. On considère les variables X_i comme indépendantes et identiquement distribués (i.i.d.) suivant une loi de Bernoulli de paramètre $\theta \in [0, 1]$ inconnus.

- Qu'est ce qu'une loi de Bernoulli ?

$\Rightarrow$ du nom du mathématicien suisse Jacques Bernoulli, est une distribution discrète de probabilité, où X prend la valeur 1 avec la probabilité $p = \theta$ et 0 avec la probabilité $q = 1 - \theta$. En d'autres termes,

$$\mathbb{P}[X = x] = \begin{cases} p & si \quad x = 1 \\ 1 - p & si \quad x = 0 \\ 0 & sinon \end{cases}$$

Exemples et problématique

L'état décide d'organiser un référendum sur la légalisation du cannabis. À cette occasion, on interroge n français. On note $X_i = 1$ si l'individu répond « oui » et $X_i = 0$ si l'individu répond « non ». Les individus sont choisis au hasard dans la population en âge de voter. On considère les variables X_i comme indépendantes et identiquement distribués (i.i.d.) suivant une loi de Bernoulli de paramètre $\theta \in [0, 1]$ inconnus.

- Qu'est ce qu'une loi de Bernoulli ?

$\Rightarrow$ du nom du mathématicien suisse Jacques Bernoulli, est une distribution discrète de probabilité, où X prend la valeur 1 avec la probabilité $p = \theta$ et 0 avec la probabilité $q = 1 - \theta$. En d'autres termes,

$$\mathbb{P}[X = x] = p^x (1 - p)^{1-x} \qquad x \in \{0, 1\}$$

Exemples et problématique

L'état décide d'organiser un référendum sur la légalisation du cannabis. À cette occasion, on interroge n français. On note $X_i = 1$ si l'individu répond « oui » et $X_i = 0$ si l'individu répond « non ». Les individus sont choisis au hasard dans la population en âge de voter. On considère les variables X_i comme indépendantes et identiquement distribués (i.i.d.) suivant une loi de Bernoulli de paramètre $\theta \in [0, 1]$ inconnus.

Ce paramètre θ est la proportion de Français qui voteraient « oui » si le référendum se déroulait le jour où le sondage a eu lieu. Au vu des réalisation des variables aléatoires X_i, on cherche à déterminer la valeur de θ et l'on désire savoir si le projet de loi va être adopté. C'est à dire que l'on :

- souhaite savoir si le projet de loi va être adopté
- tester si θ va être supérieur à $1/2$.

Exemples et problématique

Une entreprise considère que le nombre journalier d'appel téléphonique que passe chacun de ses clients suit une loi de Poisson. Cette hypothèse lui permettant de fixer ses tarifs, elle souhaite la **tester** à l'aide du décompte (sur une journée) du nombre d'appel téléphonique passés par n clients choisis au hasard dans une base de données. Si elle accepte cette hypothèse, elle souhaitera alors également **estimer** le paramètre θ de la loi.

- Quelle est le type de valeurs observées ?
- Que permet de représenter la loi de poisson ?

Exemples et problématique

Une entreprise considère que le nombre journalier d'appel téléphonique que passe chacun de ses clients suit une loi de Poisson. Cette hypothèse lui permettant de fixer ses tarifs, elle souhaite la **tester** à l'aide du décompte (sur une journée) du nombre d'appel téléphonique passés par n clients choisis au hasard dans une base de données. Si elle accepte cette hypothèse, elle souhaitera alors également **estimer** le paramètre θ de la loi.

- Quelle est le type de valeurs observées ?

$\Rightarrow$ X prend des valeurs strictement entières dans $\mathbb{N}$

Exemples et problématique

Une entreprise considère que le nombre journalier d'appel téléphonique que passe chacun de ses clients suit une loi de Poisson. Cette hypothèse lui permettant de fixer ses tarifs, elle souhaite la **tester** à l'aide du décompte (sur une journée) du nombre d'appel téléphonique passés par n clients choisis au hasard dans une base de données. Si elle accepte cette hypothèse, elle souhaitera alors également **estimer** le paramètre θ de la loi.

- Que permet de représenter la loi de Poisson ?

$\Rightarrow$ La loi de Poisson est une loi de probabilité discrète qui décrit le comportement du nombre d'événements se produisant dans un intervalle de temps fixé. Pour l'utiliser, il faut supposer que ces évènements se répète en moyenne de la même façon sur chaque intervalle de temps considérée, et indépendamment du temps écoulé depuis l'événement précédent.

Exemples et problématique

La loi de Poisson est une loi de probabilité discrète qui décrit le comportement du nombre d'événements se produisant dans un intervalle de temps fixé. Pour l'utiliser, il faut supposer que ces évènements se répète en moyenne de la même façon sur chaque intervalle de temps considérée, et indépendamment du temps écoulé depuis l'événement précédent.

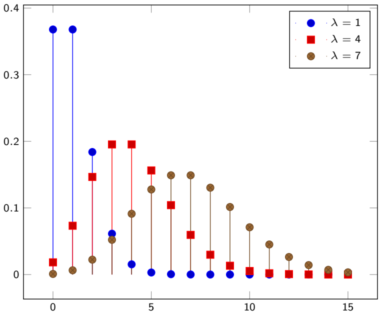
Si le nombre moyen d'évènements dans un intervalle de temps fixé est λ, alors la probabilité qu'il existe exactement k occurrences (k étant un entier naturel, $k = 0, 1, 2 \dots$) est

$$\mathbb{P}[X = k] = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}$$

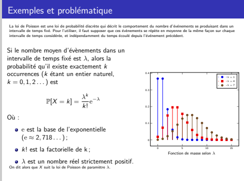
Où :

- e est la base de l'exponentielle ($\mathrm{e} \approx 2,718\dots$) ;
- $k!$ est la factorielle de k ;
- λ est un nombre réel strictement positif.

On dit alors que X suit la loi de Poisson de paramètre λ.



Fonction de masse selon λ



En théorie des probabilités, la fonction de masse est la fonction qui donne la probabilité d'un résultat élémentaire d'une expérience. Elle se distingue de la densité de probabilité en ceci que les densités de probabilité ne sont définies que pour des variables aléatoires absolument continues, et que c'est leur intégrale sur un domaine qui a valeur de probabilité (et non leurs valeurs elles-mêmes).

Exemples et problématique

La loi de Poisson est une loi de probabilité discrète qui décrit le comportement du nombre d'événements se produisant dans un intervalle de temps fixé. Pour l'utiliser, il faut supposer que ces évènements se répète en moyenne de la même façon sur chaque intervalle de temps considérée, et indépendamment du temps écoulé depuis l'événement précédent.

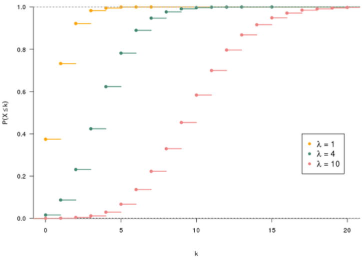
Si le nombre moyen d'évènements dans un intervalle de temps fixé est λ, alors la probabilité qu'il existe exactement k occurrences (k étant un entier naturel, $k = 0, 1, 2 \ldots$) est

$$\mathbb{P}[X = k] = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda}$$

Où :

- e est la base de l'exponentielle ($\mathrm{e} \approx 2,718\ldots$) ;
- $k!$ est la factorielle de k ;
- λ est un nombre réel strictement positif.

On dit alors que X suit la loi de Poisson de paramètre λ.



Conclusion

- L'objet de la **statistique inférentielle** est de répondre aux problèmes décris dans ces exemples.
- En **théorie des probabilité**, on suppose que la loi est connue et on souhaite caractériser le comportement d'une variable aléatoire qui suit cette loi.
- L'objectif de la **statistique** est le contraire : à partir de la connaissance de la variable, que peut-on dire de la loi de cette variable ?

Modèle statistique

POURQUOI ?

POURQUOI ?

La notion de modèle statistique nous donnera le cadre mathématique nécessaire pour la présentation rigoureuse des problèmes statistique décrit dans le paragraphe précédent et leur résolution.

Modèle statistique

- Un **modèle statistique** est la donnée d'un espace Ω mesurée par une tribu $\mathcal{A}$ et une famille $(\mathbb{P}_\theta)_{\theta \in \Theta}$ de lois de probabilité. Le modèle associé est noté $(\Omega, \mathcal{A}, \mathbb{P}_\theta, \theta \in \Theta)$. Quand il existe un $d \in \mathbb{N}^*$ tel que $\Theta \subset \mathbb{R}^d$, le modèle est dit **paramétrique**. Sinon il est dit **non paramétrique**.

$\Rightarrow$ Ω peut se voir comme toutes les valeurs possible que peuvent prendre vos observations.

$\Rightarrow$ Le terme tribu $\mathcal{A}$ veut simplement dire que vos observations sont non vides, distinguables et dénombrable.

$\Rightarrow$ θ est la valeur possible que peut prendre un paramètre (par exemple la moyenne des notes d'une classe comprise entre 0 et 20). θ est un sous espace de Θ (par exemple la moyenne de la classe est un sous espace de $\mathcal{R}$).

Modèle statistique

- Un **modèle statistique** est la donnée d'un espace Ω mesurée par une tribu $\mathcal{A}$ et une famille $(\mathbb{P}_\theta)_{\theta\in\Theta}$ de lois de probabilité. Le modèle associé est noté $(\Omega, \mathcal{A}, \mathbb{P}_\theta, \theta \in \Theta)$. Quand il existe un $d \in \mathbb{N}^*$ tel que $\Theta \subset \mathbb{R}^d$, le modèle est dit **paramétrique**. Sinon il est dit **non paramétrique**.

$\Rightarrow$ Certaines lois sont regroupées par famille par rapport à certaines propriétés de leur densité ou de leur fonction de masse.

$\Rightarrow$ Lorsque les observations "ressemblent" à des fonctions connues n'ayant besoin que de d paramètres pour représenter une densité de probabilité des observations, on dit que le modèle est paramétrique. Par exemple la réponse des français à mon référendum suit une loi de Bernoulli ne nécessitant qu'un paramètre.

$\Rightarrow$ Lorsque les observations " ne ressemblent à rien"  on dira que le modèle est non paramétrique.

Modèle statistique

Observation

- Une **observation** X est une variable aléatoire à valeur dans Ω et dont la loi appartient à la famille $(\mathbb{P}_\theta)_{\theta \in \Theta}$

$\Rightarrow$ Par exemple la réponse d'un esiearque au référendum sera 1 où 0 et suit une loi de Bernoulli de paramètre $\theta = 3/4$ tel que $\theta \in \Theta = [0, 1]$ 

Modèle statistique

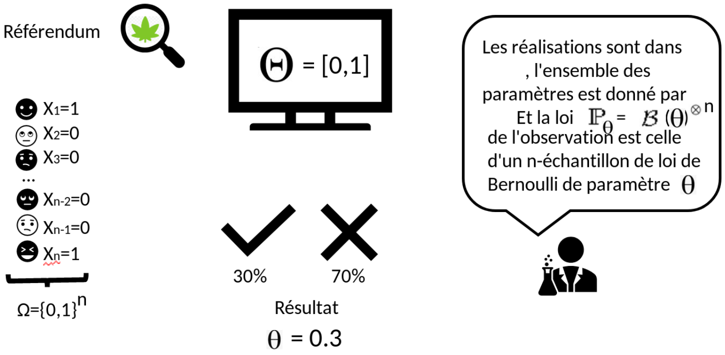
n-échantillon

- Pour $n \in \mathbb{N}^*$, un ***n*-échantillon** de loi ν est la donnée de n variables $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon la loi ν.

$\Rightarrow$ On dit généralement que X prend (par exemple) des valeurs dans $\{0, 1\}$. Si l'on veut être plus pénible  on dira que les réalisations sont dans $\Omega = \{0, 1\}^n$.

Conclusion

L'état décide d'organiser un référendum sur la légalisation du cannabis. À cette occasion, on interroge n français. On note $X_i = 1$ si l'individu répond « oui » et $X_i = 0$ si l'individu répond « non ». Les individus sont choisis au hasard dans la population en âge de voter. On considère les variables X_i comme indépendantes et identiquement distribués (i.i.d.) suivant une loi de Bernoulli de paramètre $\theta \in [0, 1]$ inconnu.



Quizz

Quizz

La loi de Bernoulli permet de :

- modéliser des variables aléatoires pouvant prendre au moins deux valeurs (modalités)
- modéliser des variables aléatoires pouvant prendre au maximum deux valeurs (modalités)
- modéliser des variables aléatoires pouvant prendre deux valeurs (modalités)
- modéliser des variables aléatoires pouvant prendre plus de deux valeurs (modalités)

Quizz

La loi de Bernoulli permet de :

- modéliser des variables aléatoires pouvant prendre au moins deux valeurs (modalités)
- modéliser des variables aléatoires pouvant prendre au maximum deux valeurs (modalités)
- modéliser des variables aléatoires pouvant prendre deux valeurs (modalités)
- modéliser des variables aléatoires pouvant prendre plus de deux valeurs (modalités)

Quizz

La loi de Poisson permet de :

- décrire le comportement d'un nombre d'évènements produits pendant une période fixée.
- décrire la valeur d'évènements produite pendant une période fixée.
- décrire le prix d'un poisson pendant une période de solde
- décrire une somme de valeurs produites pendant une période fixée.

Quizz

La loi de Poisson permet de :

- décrire le comportements d'un nombre d'évènements produits pendant une période fixée.
- décrire la valeur d'évènements produite pendant une période fixée.
- décrire le prix d'un poisson pendant une période de solde
- décrire une somme de valeurs produites pendant une période fixée.

Quizz

Si l'on veut représenter la fonction de probabilité d'une variable discrète il faudra utiliser une :

- fonction de densité de probabilité
- fonction aléatoire
- fonction analytique complexe zêta de Riemann
- fonction de masse

Quizz

Si l'on veut représenter la fonction de probabilité d'une variable discrète il faudra utiliser une :

- fonction de densité de probabilité
- fonction aléatoire
- fonction analytique complexe zêta de Riemann
- fonction de masse

Quizz

L'objectif de la statistique est de :

- trouver le paramètre d'une loi
- dire si les observations d'une variable suivent une certaine loi.
- prédire la prochaine valeur de X après n observations
- prendre la tête.

Quizz

L'objectif de la statistique est de :

- trouver le paramètre d'une loi
- dire si les observations d'une variable suivent une certaine loi.
- prédire la prochaine valeur de X après n observations.
- prendre la tête.

Quizz

Un modèle statistique paramétrique est noté :

- $(\Omega, \mathcal{A}, \mathbb{P}_\theta, \theta \in \Theta)$
- $(\Sigma, \mathcal{A}, \mathbb{P}_\theta, \theta \in \Theta)$
- $(\Omega, \mathcal{B}, \mathbb{P}_\theta, \theta \in \Theta)$
- $(\Omega, \mathcal{A}, \mathbb{P}_\omega, \omega \in \Omega)$

Quizz

Un modèle statistique paramétrique est noté :

- $(\Omega, \mathcal{A}, \mathbb{P}_\theta, \theta \in \Theta)$
- $(\Sigma, \mathcal{A}, \mathbb{P}_\theta, \theta \in \Theta)$
- $(\Omega, \mathcal{B}, \mathbb{P}_\theta, \theta \in \Theta)$
- $(\Omega, \mathcal{A}, \mathbb{P}_\omega, \omega \in \Omega)$

Quizz

Une entreprise considère que le nombre journalier d'appel téléphonique que passe chacun de ses clients suit une loi de Poisson. Cette hypothèse lui permettant de fixer ses tarifs, elle souhaite la **tester** à l'aide du décompte (sur une journée) du nombre d'appel téléphonique passés par n clients choisis au hasard dans une base de données. Si elle accepte cette hypothèse, elle souhaitera alors également **estimer** le paramètre θ de la loi. $\Omega = \mathbb{N}^n$?

- Vrais
- Faux

Quizz

Une entreprise considère que le nombre journalier d'appel téléphonique que passe chacun de ses clients suit une loi de Poisson. Cette hypothèse lui permettant de fixer ses tarifs, elle souhaite la **tester** à l'aide du décompte (sur une journée) du nombre d'appel téléphonique passés par n clients choisis au hasard dans une base de données. Si elle accepte cette hypothèse, elle souhaitera alors également **estimer** le paramètre θ de la loi.
$\Omega = \mathbb{N}^n$?

- Vrais
- Faux

Quizz

Une entreprise considère que le nombre journalier d'appel téléphonique que passe chacun de ses clients suit une loi de Poisson. Cette hypothèse lui permettant de fixer ses tarifs, elle souhaite la **tester** à l'aide du décompte (sur une journée) du nombre d'appel téléphonique passés par n clients choisis au hasard dans une base de données. Si elle accepte cette hypothèse, elle souhaitera alors également **estimer** le paramètre θ de la loi. On défini Θ tel que pour tout $\Theta = \left\{\theta = (p_0, p_1, ...) : p_i \geq 0, \sum_{i=0}^{+\infty} p_i = 1\right\}$ et $\mathbb{P}_\theta = (\nu_\theta)^{\otimes n}$ la loi de l'observation, où (μ_θ) est la loi de probabilité paramétrée par θ Les p_i représentent :

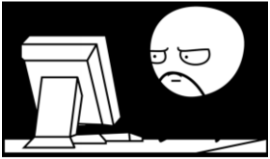
- le nombre journalier d'appel
- la probabilité d'obtenir un nombre i journalier d'appel

Quizz

Une entreprise considère que le nombre journalier d'appel téléphonique que passe chacun de ses clients suit une loi de Poisson. Cette hypothèse lui permettant de fixer ses tarifs, elle souhaite la **tester** à l'aide du décompte (sur une journée) du nombre d'appel téléphonique passés par n clients choisis au hasard dans une base de données. Si elle accepte cette hypothèse, elle souhaitera alors également **estimer** le paramètre θ de la loi. On défini Θ tel que pour tout $\Theta = \left\{\theta = (p_0, p_1, ...) : p_i \geq 0, \sum_{i=0}^{+\infty} p_i = 1\right\}$ et $\mathbb{P}_\theta = (\nu_\theta)^{\otimes n}$ la loi de l'observation, où (μ_θ) est la loi de probabilité paramétrée par θ Les p_i représentent :

- le nombre journalier d'appel
- la probabilité d'obtenir un nombre i journalier d'appel

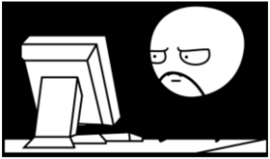
TD n° 0 : Installation de R (10/20 minutes)



Pensez à sauvegarder vos commandes !

TD n° 1

TD n° 1 : Distribution, densité de probabilité et tests (1h30)



- Observation sur les données
- Hypothèse sur la loi de distribution

Pensez à sauvegarder vos commandes !

For Further Reading I

R. Rivoirard et G. Stoltz
Statistique en action .
Vuibert, 2006.