

Statistiques (STA1)

Cours I – Rappels de statistiques : estimation, tests et intervalles de confiance

Luca Ganassali

Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay

Jeudi 18 septembre 2025

Introduction

On dispose d'un échantillon de 6 relevés du temps de trajet (en min) domicile/bureau d'un employé : $x = (15, 17, 15, 18, 16, 15)$.

On souhaite répondre aux questions suivantes :

1. Quelle est la durée moyenne estimée d'un trajet sur l'année?
2. Peut-on affirmer avec peu de risque que la durée moyenne d'un trajet est supérieure à 15 min?
3. Peut-on donner un intervalle dans lequel la durée moyenne se trouve ?
4. Si d'autres facteurs sont disponibles pour chaque relevé en plus du temps de trajet (météo, jour de la semaine, horaire...), certains d'entre eux influencent-ils la durée de trajet?
5. Quelle sera la durée d'un trajet demain?

- En **probabilités**, on cherche à étudier les propriétés d'une variable aléatoire qui suit une loi \mathbb{P} connue :

$$\mathbb{P} \longrightarrow \text{propriétés de } X \sim \mathbb{P}$$

- En **statistiques**, à partir d'observations d'une loi \mathbb{P} inconnue, on cherche à apprendre (inférer) des propriétés de cette loi pour répondre à une question :

$$\text{observations } x \text{ d'un } X \sim \mathbb{P} \longrightarrow \text{propriétés de } \mathbb{P}$$

- **Modélisation** : choisir convenablement un ensemble de lois \mathbb{P} possibles, en adéquation avec (i) nos connaissances préalables, (ii) nos objectifs, (iii) les données, (iv) les capacités de calcul...
- **Estimation** : estimer la valeur d'un paramètre d'intérêt de la loi \mathbb{P}
- **Intervalle de confiance** : savoir encadrer la valeur du paramètre d'intérêt entre deux bornes
- **Test** : prendre une décision sur une hypothèse à l'aide des données
- **Prédiction** : pouvoir prédire la valeur prise par une nouvelle variable non encore observée

Enjeux:

- Comprendre comment construire, comparer, choisir des procédures
- Quantifier la fiabilité, le risque de l'information obtenue

On dispose d'un échantillon de 6 relevés du temps de trajet (en min) domicile/bureau d'un employé : $x = (15, 17, 15, 18, 16, 15)$.

On souhaite répondre aux questions suivantes :

1. Quelle est la durée moyenne estimée d'un trajet sur l'année?
(estimation)
2. Peut-on affirmer avec peu de risque que la durée moyenne d'un trajet est supérieure à 15 min? (test)
3. Peut-on donner un intervalle dans lequel la durée moyenne se trouve ?
(intervalle de confiance)
4. Si d'autres facteurs sont disponibles pour chaque relevé en plus du temps de trajet (météo, jour de la semaine, horaire...), certains d'entre eux influencent-ils la durée de trajet? (modélisation)
5. Quelle sera la durée d'un trajet demain? (prédiction)

- Cours I (aujourd'hui) – Rappels de statistiques : estimation, tests et intervalles de confiance
- Cours II – Information de Fisher, estimation par maximum de vraisemblance
- Cours III – Tests uniformément plus puissants, Théorème de Neyman-Pearson
- Cours IV – Vecteurs gaussiens. Modèles linéaire et linéaire gaussien
- Cours V – Régression linéaire 1/2 : Estimateur des moindres carrés, propriétés générales, tests de Student
- Cours VI – Régression linéaire 2/2 : intervalle de confiance pour un coefficient, interprétation géométrique, erreur de prédiction, modèles emboîtés, lecture de résultats sur R...

Être capable, en utilisant les bases de la statistique mathématique, de :

- Définir une modélisation adaptée à un jeu de données
- Construire des estimateurs, en étudier le risque, l'efficacité et l'asymptotique dans des cas simples
- Construire des tests, des intervalles de confiance
- Travailler avec le modèle linéaire: estimation, prédiction, interprétation géométrique.
- Fitter un modèle linéaire avec un logiciel et interpréter les résultats obtenus
- Prendre en compte le risque de toute décision statistique et le quantifier.

Documents de support de cours : Moodle/E-Campus

Evaluation : un examen le vendredi 14 novembre 2025, 10h-12h, qui pourra comporter des questions théoriques et des questions pratiques d'interprétation de résultats. Il n'y aura pas de code informatique à écrire.

Chargés de TDs : G. Debaussart, A. Janon, H. Henneuse, B. Bouriquet, J. Capitaio, L. Ganassali

Modélisation statistique

Modéliser l'expérience, c'est proposer un ensemble de lois théoriques pour la distribution possible des données observées z (échantillon).

- Un **modèle** est la donnée de $\mathcal{M} = (\mathcal{Z}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ avec
 - $(\mathcal{Z}, \mathcal{A})$ espace mesurable (on omettra \mathcal{A} dans la suite)
 - $(\mathbb{P}_\theta)_{\theta \in \Theta}$ une famille de lois de probabilité sur \mathcal{Z} , indexées par un ensemble Θ .

Quand $\Theta \subset \mathbb{R}^d$, le modèle est dit **paramétrique**.

- Les **données** z sont une réalisation (valeur) particulière prises par la variable aléatoire Z dans \mathcal{Z} , et dont la loi appartient au modèle.
- Lorsque Z est de la forme $Z = (X_1, \dots, X_n)$ avec les X_i indépendantes et identiquement distribuées (i.i.d.), on dit que X est un **échantillon** de taille n , ou n -échantillon. Dans ce cas, \mathcal{Z} s'écrit $\mathcal{Z} = \mathcal{X}^n$ et \mathbb{P}_θ s'écrit $\mathbb{P}_\theta = (\eta_\theta)^{\otimes n}$.
- Le modèle \mathcal{M} est dit **identifiable** lorsque pour deux paramètres différents, la loi de Z est différente, i.e. $\theta \in \Theta \mapsto \mathbb{P}_\theta$ est injective.

Exemples de modèles simples

- Étude de la moyenne (espérance) d'un temps de trajet:

$$\mathcal{M} = (\mathbb{R}^n, ((\mathcal{N}(\mu, \sigma^2))^{\otimes n})_{(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+}).$$

- Variance supposée connue, par ex. $\sigma = 5$: on modélise seulement la moyenne inconnue, $\theta = \mu$. Le processus de décision statistique d'après les données peut utiliser la valeur connue $\sigma = 5$.
 - Variance inconnue: fait partie du modèle, $\theta = (\mu, \sigma^2)$. Le processus de décision statistique d'après les données ne **peut pas** dépendre de σ (ni de μ).
- Estimation d'une proportion par sondage $\mathcal{M} = (\{0, 1\}^n, (B(p))^{\otimes n})_{p \in [0, 1]}$
 - Comparaison du rendement de maïs sous deux conditions de culture :

$$(X_1, \dots, X_n) \sim \mathcal{N}(\mu_x, \sigma_x^2)^{\otimes n} \text{ et } (Y_1, \dots, Y_m) \sim \mathcal{N}(\mu_y, \sigma_y^2)^{\otimes m} \text{ indépendants}$$

(Les observations sont indépendantes mais pas de même loi)

Tous ces modèles sont identifiables. Exemple de modèles non identifiables : cf TD.

Estimation ponctuelle

Soit $\theta \in \Theta$ paramètre d'une loi \mathbb{P}_θ et $Z = (X_1, \dots, X_n)$ un n -échantillon de cette loi.

Une **statistique** est une **variable aléatoire** T_n , fonction mesurable de l'échantillon et calculable à partir de l'échantillon :

$$T_n = F(X_1, \dots, X_n).$$

Un **estimateur** est une statistique utilisée pour estimer un paramètre ou une quantité d'intérêt $\varphi(\theta) \in \mathbb{R}^k$.

On notera $T_n = \hat{\varphi}$.

Exemple : si Z est un n -échantillon de variables d'espérance finie μ , un estimateur de μ est par exemple $\hat{\mu} = \bar{X} := \frac{1}{n} \sum_{i=1}^n X_i$, appelée moyenne empirique.

Biais, variance, risque quadratique d'un estimateur

Soit $\hat{\varphi}$ un estimateur de $\varphi(\theta) \in \mathbb{R}^k$, fonction du paramètre d'une loi \mathbb{P}_θ .

- On appelle (fonction de) **biais** de $\hat{\varphi}$ pour $\varphi(\theta)$ le vecteur

$$b_\theta(\hat{\varphi}) = \mathbb{E}_\theta[\hat{\varphi}] - \varphi(\theta) \in \mathbb{R}^k,$$

qui est **fonction du vrai paramètre** θ . Si $b_\theta(\hat{\varphi}) = \mathbf{0}_k$ pour tout $\theta \in \Theta$, $\hat{\varphi}$ est dit **sans biais** pour estimer $\varphi(\theta)$.

- On appelle **matrice de covariance** de $\hat{\varphi}$ la valeur (**fonction du paramètre** θ)

$$\text{Var}_\theta(\hat{\varphi}) = \mathbb{E}_\theta[(\hat{\varphi} - \mathbb{E}_\theta[\hat{\varphi}])(\hat{\varphi} - \mathbb{E}_\theta[\hat{\varphi}])^T] \in \mathbb{R}^{k \times k}$$

- On appelle **risque quadratique** de $\hat{\varphi}$ la valeur (**fonction du paramètre** θ)

$$R_\theta(\hat{\varphi}) = \mathbb{E}_\theta[\|\hat{\varphi} - \varphi(\theta)\|^2] \in \mathbb{R}_+,$$

où $\|\cdot\|$ est la norme euclidienne canonique sur \mathbb{R}^k .

Proposition (Pythagore du statisticien)

On a, pour tout $\theta \in \Theta$,

$$R_\theta(\hat{\varphi}) = \|\mathbf{b}_\theta(\hat{\varphi})\|^2 + \text{Tr}(\text{Var}_\theta(\hat{\varphi})).$$

Preuve.

Il suffit d'écrire, pour tout $\theta \in \Theta$,

$$\begin{aligned} R_\theta(\hat{\varphi}) &= \mathbb{E}_\theta[\|\hat{\varphi} - \mathbb{E}_\theta[\hat{\varphi}] + \mathbb{E}_\theta[\hat{\varphi}] - \varphi(\theta)\|^2] \\ &= \mathbb{E}_\theta[\|\hat{\varphi} - \mathbb{E}_\theta[\hat{\varphi}]\|^2] + \mathbb{E}_\theta[\|\mathbb{E}_\theta[\hat{\varphi}] - \varphi(\theta)\|^2] + 2(\mathbb{E}_\theta[\hat{\varphi}] - \varphi(\theta))^T \mathbb{E}_\theta[\hat{\varphi} - \mathbb{E}_\theta[\hat{\varphi}]] \\ &= \mathbb{E}_\theta[\|\hat{\varphi} - \mathbb{E}_\theta[\hat{\varphi}]\|^2] + \|\mathbf{b}_\theta(\hat{\varphi})\|^2, \end{aligned}$$

et par cyclicité et linéarité de la trace, et linéarité de l'espérance,

$$\begin{aligned} \mathbb{E}_\theta[\|\hat{\varphi} - \mathbb{E}_\theta[\hat{\varphi}]\|^2] &= \mathbb{E}_\theta[\text{Tr}[(\hat{\varphi} - \mathbb{E}_\theta[\hat{\varphi}])^T (\hat{\varphi} - \mathbb{E}_\theta[\hat{\varphi}])]] \\ &= \mathbb{E}_\theta[\text{Tr}[(\hat{\varphi} - \mathbb{E}_\theta[\hat{\varphi}]) (\hat{\varphi} - \mathbb{E}_\theta[\hat{\varphi}])^T]] \\ &= \text{Tr}[\mathbb{E}_\theta[(\hat{\varphi} - \mathbb{E}_\theta[\hat{\varphi}]) (\hat{\varphi} - \mathbb{E}_\theta[\hat{\varphi}])^T]] \\ &= \text{Tr}(\text{Var}_\theta(\hat{\varphi})). \end{aligned}$$

Un estimateur δ_1 de $\varphi(\theta)$ **domine** l'estimateur δ_2 si, pour tout $\theta \in \Theta$,

$$R_\theta(\delta_1) \leq R_\theta(\delta_2),$$

cette inégalité étant stricte pour au moins une valeur de θ .

Il n'existe en général pas d'estimateur dominant tous les autres.

Dans le cas d'un modèle à échantillon i.i.d. de taille n , pour $\hat{\varphi} = \hat{\varphi}_n$ un estimateur de $\varphi(\theta)$, on dira que $\hat{\varphi}$ est **fortement consistant** si il converge presque sûrement vers $\varphi(\theta)$, i.e.

$$\forall \theta \in \Theta, \mathbb{P}_\theta(\lim_{n \rightarrow \infty} \hat{\varphi}_n = \varphi(\theta)) = 1$$

Tests

- Un constructeur automobile annonce une consommation $\mu_0 = 6.32\ell/100\text{ km}$, avec un écart type $\sigma = 0.21\ell/100\text{ km}$, pour des véhicules d'un type donné.
- Un organisme indépendant suspecte une sous-estimation de cette consommation et indique que la consommation s'élèverait à $\mu_1 = 6.45\ell/100\text{ km}$. Il ne remet pas en cause la variance.
- Sur un 30-échantillon, on observe $\bar{x} = 6.43\ell/100\text{ km}$.

Doit-on incriminer le constructeur?

1. **Définir un modèle** : Les individus sont les véhicules, $i = 1, \dots, n$, la variable observée X_i est la consommation. On suppose les $X_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d., μ inconnue et variance connue $\sigma^2 = (0.21)^2$
2. **Définir les hypothèses**
 - H_0 conso. conforme au constructeur: $\mu = \mu_0 = 6.32$
 - H_1 conso. suspectée par l'organisme: $\mu = \mu_1 = 6.45$

Choisir, à partir de l'observation du n -échantillon, entre les deux hypothèses H_0 et H_1 , tel que le risque de *choisir H_1 alors que H_0 est vrai* soit faible et maîtrisé. C'est le **risque de première espèce** α (5%, 10%, ...) aussi appelé **niveau**.

3. Définir une statistique de test

- Travailler à partir de \bar{X} , moyenne des consommations de $n = 30$ véhicules
- Loi sous H_0 ,

$$\bar{X} \sim \mathcal{N}(\mu_0, \sigma^2/n) \text{ donc } T = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} \sim \mathcal{N}(0, 1)$$

4. Définir la région de rejet

- Choisir a priori un niveau α , calibrant la probabilité de rejet de H_0 à tort ($\alpha = 5\%$ par exemple)
- Calculer le **seuil** pour la région de rejet

$$\begin{aligned} \alpha &= \mathbb{P}_{H_0}(\text{rejeter } H_0) \\ &= \underbrace{\mathbb{P}_{H_0}(T > q_{1-\alpha}^*)}_{\mathcal{R} =]q_{1-\alpha}^*; \infty[, \text{ Région de rejet pour } T} \\ &= \underbrace{\mathbb{P}_{H_0}(\bar{X} > \mu_0 + q_{1-\alpha}^* \frac{\sigma}{\sqrt{n}})}_{\mathcal{R} =]\mu_0 + q_{1-\alpha}^* \frac{\sigma}{\sqrt{n}}; \infty[, \text{ Région de rejet pour } \bar{x}} \end{aligned}$$

avec $q_{1-\alpha}^*$ le quantile d'une loi $\mathcal{N}(0, 1)$ d'ordre $1 - \alpha$

5. Décider:

- si T est dans la région de rejet, on **rejette** H_0
- sinon, on **ne rejette pas** H_0 et on la conserve, faute de preuves suffisantes

Ici, $\bar{x} = 6.43\ell/100\text{ km}$,

$$t_{obs} = \frac{6.43 - 6.32}{0.21/\sqrt{30}} = 2.86 > 1.64$$

au niveau $\alpha = 5\%$, les données sont significatives pour rejeter H_0 , le constructeur a minimisé la consommation, avec un risque (de première espèce) α .

Un autre cas de figure:

- si la même consommation a été observée sur un échantillon de $n = 9$ véhicules

$$t_{obs} = \frac{6.43 - 6.32}{0.21/\sqrt{9}} = 1.57 < 1.64$$

on ne peut pas rejeter le fait que le constructeur a sous-estimé la consommation, on ne peut rejeter H_0 qu'on accepte par défaut

- avec quelle erreur?

Une autre façon de se tromper: l'**erreur de seconde espèce**: ne pas rejeter H_0 alors que H_1 est vraie.

Sous H_1 , $\bar{X} \sim \mathcal{N}(\mu_1, \frac{\sigma^2}{n})$ et le risque de seconde espèce est

$$\begin{aligned}\beta &= \mathbb{P}_{H_1}(\text{conserver } H_0) \\ &= \mathbb{P}_{H_1}(\bar{X} < \mu_0 + q_{1-\alpha}^* \frac{\sigma}{\sqrt{n}}) \\ &= F^* \left(\sqrt{n} \frac{\mu_0 - \mu_1}{\sigma} + q_{1-\alpha}^* \right)\end{aligned}$$

A.N.: $n = 9$, $\beta \simeq 0.41$. La **puissance** du test $\pi = 1 - \beta$ n'est pas très grande.

Un **test** est une procédure de décision qui permet de trancher, au vu des résultats d'un échantillon, entre deux hypothèses l'**hypothèse nulle** H_0 et une hypothèse **alternative** H_1 , dont une seule est vraie.

La **région critique** ou région de **rejet** \mathcal{R} est l'ensemble des valeurs de la statistique de test T qui conduisent à écarter H_0 au profit de H_1 .

Le **niveau** du test est le **risque de première espèce** $\alpha = \mathbb{P}_{H_0}(T \in \mathcal{R})$.

La **puissance** du test est $\pi = 1 - \beta$ où $\beta = \mathbb{P}_{H_1}(T \notin \mathcal{R})$ est le **risque de seconde espèce**.

A l'issue du test, les quatre situations suivantes sont possibles

	Choix H_0	Choix H_1
H_0 vraie	$1 - \alpha$ bonne décision	$\alpha = \mathbb{P}_{H_0}(T \in \mathcal{R})$ risque première espèce mauvaise décision
H_1 vraie	$\beta = \mathbb{P}_{H_1}(T \notin \mathcal{R})$ risque seconde espèce mauvaise décision	$\pi = 1 - \beta$ puissance bonne décision

- Définir le **modèle**
- Définir les **hypothèses nulle** H_0 et **alternative** H_1
- Choisir une **statistique de test** $T(Z)$, calculer sa **loi sous** H_0 . On fera en sorte que la loi de $T(Z)$ sous H_0 ne dépende plus d'aucun paramètre du modèle : c'est une variable **pivotal**.
- Définir la **règle de décision** en calibrant la région de rejet \mathcal{R} suivant le risque α . Le test prend alors la forme :

$$\mathbf{1}_{\{T(Z) \in \mathcal{R}\}} \cdot$$

- Calcul de la statistique observée et **décision**: rejet ou acceptation de H_0 .

- Le risque n'est contrôlé que pour H_0
 - La véritable décision est celle qui rejette H_0 .
 - H_0 et H_1 ne sont pas interchangeables.
- Il faut connaître la loi de la statistique de test sous H_0
- Il faut que cette loi soit différente sous H_1
- Entre deux tests de même risque de 1ère espèce α , il faut choisir le plus puissant
 - La région de rejet de la forme $\{T < q_{0.05}^*\}$ est aussi de risque 5%, mais elle n'a aucune puissance pour détecter le cas $\mu_1 > \mu_0$

La valeur observée n'a pas servi à construire la région de rejet qui a été définie a priori en fonction de la problématique fixée.

- Ainsi, pour tester

$$H_0 : \theta = \theta_0 \text{ contre } H_1 : \theta > \theta_0$$

on utilise la même région de rejet que pour le test d'hypothèse simple

$$H_0 : \theta = \theta_0 \text{ contre } H_1 : \theta = \theta_1 > \theta_0$$

- mais la **puissance** devient une **fonction** de θ :

$$\theta_1 \in \Theta_1 = \{\theta | \theta > \theta_0\}, \pi(\theta_1) = \mathbb{P}_{\theta_1}(T \in \mathcal{R}) = 1 - \beta(\theta_1).$$

- Hypothèses simples

$$H_0 : \theta = \theta_0 \text{ contre } H_1 : \theta = \theta_1$$

- Test unilatéral pour une hypothèse nulle composite

$$H_0 : \theta \leq \theta_0 \text{ contre } H_1 : \theta > \theta_0$$

- Test unilatéral pour une hypothèse nulle composite

$$H_0 : \theta \geq \theta_0 \text{ contre } H_1 : \theta < \theta_0$$

- Test bilatéral pour une hypothèse nulle simple

$$H_0 : \theta = \theta_0 \text{ contre } H_1 : \theta \neq \theta_0$$

- De façon générale:

$$H_0 : \theta \in \Theta_0 \text{ contre } H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$$

Régions de confiance

Contexte : on cherche à quantifier l'incertitude que l'on a au vu des données sur la valeur d'un paramètre d'intérêt $\varphi(\theta) \in \mathbb{R}^k$. Un ensemble aléatoire $C(Z) \subseteq \mathbb{R}^k$ est une **région de confiance** de niveau $1 - \alpha$ pour $\varphi(\theta)$ si :

$$\forall \theta \in \Theta, \quad \mathbb{P}_\theta(\varphi(\theta) \in C(Z)) \geq 1 - \alpha.$$

Dans le cas d'un n -échantillon un ensemble aléatoire $C_n(Z) \subseteq \mathbb{R}^k$ est une **région de confiance asymptotique** de niveau $1 - \alpha$ pour si :

$$\forall \theta \in \Theta, \quad \liminf_n \mathbb{P}_\theta(\varphi(\theta) \in C_n(Z)) \geq 1 - \alpha.$$

Intervalle de confiance : cas $k = 1$.

Méthode pivotale : construire $C(Z)$ à partir d'une statistique pivotale dont la loi ne dépend pas de θ .

Proposition

Soit $C(Z)$ une région de confiance pour $\varphi(\theta)$ de niveau $1 - \alpha$. Alors, pour tout t , le test rejetant H_0 si et seulement si $t \notin C(Z)$ est de niveau α pour $H_0 : \varphi(\theta) = t$ (et H_1 quelconque).

Preuve.

Par définition, $\mathbb{P}_{H_0}(t \notin C(Z)) = \mathbb{P}_{H_0}(\varphi(\theta) \notin C(Z)) \leq \alpha$.

□

Proposition

Réciproquement, pour tout t , soit φ_t un test de niveau α pour $H_0 : \varphi(\theta) = t$ (et H_1 quelconque). Alors

$$C(Z) := \{t : \varphi_t(Z) = 0\}$$

définit une région de confiance pour $\varphi(\theta)$ de niveau $1 - \alpha$.

Preuve.

Par définition,

$$\mathbb{P}(\varphi(\theta) \in C(Z)) = \mathbb{P}(\varphi_{\varphi(\theta)}(Z) = 0) = \mathbb{P}_{H_0}(\text{conserver } H_0) \geq 1 - \alpha. \quad \square$$

Loi forte des grands nombres (LGN) : Soient X_1, \dots, X_n i.i.d. d'espérance $\mu = \mathbb{E}[X_1]$ finie. Alors

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{p.s.} \mu.$$

Théorème central limite (TCL) : Soient X_1, \dots, X_n i.i.d. avec espérance finie μ et variance finie σ^2 . Alors

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Applications : ces résultats permettent de prouver la **consistance** d'un estimateur, et de construire des **intervalles de confiance asymptotiques**.

Merci !

Rdv en TD pour les questions et la pratique de
ce qui vient d'être (re)vu.