Analyse de données

Pr: MAJDOUB Soufyane

Université Privé de Fès
soufyane.majdoub@usmba.ac.ma

16 décembre 2025

Plan

1. Introduction à l'analyse inférencielle
2. Rappel mathématique : Loi normale, Khi-deux, Student, Fisher
3. Estimation par intervalle de confiance
4. Tests d'hypothèses

Plan

1. Introduction à l'analyse inférencielle
2. Rappel mathématique : Loi normale, Khi-deux, Student, Fisher
3. Estimation par intervalle de confiance
4. Tests d'hypothèses

Objectif

L'inférence statistique permet de **tirer des conclusions sur une population entière** à partir d'un **échantillon limité de données**.

Objectif

L'inférence statistique permet de **tirer des conclusions sur une population entière** à partir d'un **échantillon limité de données**.

- En pratique, il est souvent **impossible d'étudier toute la population**.
- On utilise donc un **échantillon représentatif** pour **estimer**, **comparer** ou **tester** des hypothèses.
- Exemple : plutôt que d'interroger tous les électeurs, on sonde un groupe restreint.

Statistique descriptive vs inférentielle

Deux approches complémentaires

Statistique descriptive

- Résume et décrit les données disponibles.
- Utilise des mesures : moyenne, médiane, écart-type, graphiques.
- Exemple : le salaire moyen d'un échantillon d'employés est de 7000 DH.

Statistique descriptive vs inférentielle

Deux approches complémentaires

Statistique descriptive

- Résume et décrit les données disponibles.
- Utilise des mesures : moyenne, médiane, écart-type, graphiques.
- Exemple : le salaire moyen d'un échantillon d'employés est de 7000 DH.

Statistique inférentielle

- Va au-delà de la description : **généralise à la population**.
- Utilise l'échantillon pour **estimer des paramètres** (moyenne, proportion, variance).
- S'appuie sur la **théorie des probabilités**.

Statistique descriptive vs inférentielle

Deux approches complémentaires

Statistique descriptive

- Résume et décrit les données disponibles.
- Utilise des mesures : moyenne, médiane, écart-type, graphiques.
- Exemple : le salaire moyen d'un échantillon d'employés est de 7000 DH.

Statistique inférentielle

- Va au-delà de la description : **généralise à la population**.
- Utilise l'échantillon pour **estimer des paramètres** (moyenne, proportion, variance).
- S'appuie sur la **théorie des probabilités**.

À retenir

La statistique descriptive décrit **ce que l'on voit**. L'inférence statistique prédit **ce que l'on ne voit pas**.

Idée intuitive

Du petit (échantillon) au grand (population)

Principe fondamental

Observer une partie pour comprendre le tout.

Idée intuitive

Du petit (échantillon) au grand (population)

Principe fondamental

Observer une partie pour comprendre le tout.

- On collecte un **échantillon** de taille n.
- On calcule des **statistiques** (moyenne, proportion...).
- On utilise ces résultats pour **estimer les paramètres** de la population.

Idée intuitive

Du petit (échantillon) au grand (population)

Principe fondamental

Observer une partie pour comprendre le tout.

- On collecte un **échantillon** de taille n.
- On calcule des **statistiques** (moyenne, proportion...).
- On utilise ces résultats pour **estimer les paramètres** de la population.

Exemple

Un sondage réalisé auprès de 1000 personnes sur un million d'électeurs $\rightarrow$ permet d'estimer la proportion d'intentions de vote pour un candidat.

Idée intuitive

Du petit (échantillon) au grand (population)

Principe fondamental

Observer une partie pour comprendre le tout.

- On collecte un **échantillon** de taille n.
- On calcule des **statistiques** (moyenne, proportion...).
- On utilise ces résultats pour **estimer les paramètres** de la population.

Exemple

Un sondage réalisé auprès de 1000 personnes sur un million d'électeurs $\rightarrow$ permet d'estimer la proportion d'intentions de vote pour un candidat.

Hypothèse clé

L'échantillon doit être **aléatoire et représentatif** de la population.

Exemple : sondage électoral

Application concrète de l'inférence

- Population cible : tous les électeurs d'un pays.
- Échantillon : 1200 personnes choisies aléatoirement.
- Résultat du sondage : 58% déclarent voter pour le candidat A.

Exemple : sondage électoral

Application concrète de l'inférence

- Population cible : tous les électeurs d'un pays.
- Échantillon : 1200 personnes choisies aléatoirement.
- Résultat du sondage : 58% déclarent voter pour le candidat A.

Question d'inférence

Peut-on conclure que **le candidat A est réellement favori dans la population entière** ?

Exemple : sondage électoral

Application concrète de l'inférence

- Population cible : tous les électeurs d'un pays.
- Échantillon : 1200 personnes choisies aléatoirement.
- Résultat du sondage : 58% déclarent voter pour le candidat A.

Question d'inférence

Peut-on conclure que **le candidat A est réellement favori dans la population entière** ?

- On estime la proportion p dans la population.
- On calcule un **intervalle de confiance** autour de l'estimation.
- On évalue le **risque d'erreur** lié à l'échantillonnage.

Exemple : sondage électoral

Application concrète de l'inférence

- Population cible : tous les électeurs d'un pays.
- Échantillon : 1200 personnes choisies aléatoirement.
- Résultat du sondage : 58% déclarent voter pour le candidat A.

Question d'inférence

Peut-on conclure que **le candidat A est réellement favori dans la population entière** ?

- On estime la proportion p dans la population.
- On calcule un **intervalle de confiance** autour de l'estimation.
- On évalue le **risque d'erreur** lié à l'échantillonnage.

Conclusion

L'inférence statistique permet de **transformer des données d'échantillon en connaissances généralisables**.

Plan

1. Introduction à l'analyse inférencielle
2. Rappel mathématique : Loi normale, Khi-deux, Student, Fisher
3. Estimation par intervalle de confiance
4. Tests d'hypothèses

Population

L'ensemble complet des individus, objets ou observations sur lesquels on souhaite tirer des conclusions.

Population

L'ensemble complet des individus, objets ou observations sur lesquels on souhaite tirer des conclusions.

Échantillon

Un **sous-ensemble** de la population, choisi pour représenter cette dernière. *Il sert de base à l'analyse statistique.*

Population, échantillon, statistiques et estimateurs

Notions fondamentales

Population

L'ensemble complet des individus, objets ou observations sur lesquels on souhaite tirer des conclusions.

Échantillon

Un **sous-ensemble** de la population, choisi pour représenter cette dernière. *Il sert de base à l'analyse statistique.*

Statistiques et estimateurs

- Une **statistique** est une valeur calculée à partir de l'échantillon (ex : moyenne, variance).
- Un **estimateur** est une formule utilisée pour **estimer un paramètre inconnu** de la population.

Population, échantillon, statistiques et estimateurs

Notions fondamentales

Population

L'ensemble complet des individus, objets ou observations sur lesquels on souhaite tirer des conclusions.

Échantillon

Un **sous-ensemble** de la population, choisi pour représenter cette dernière. *Il sert de base à l'analyse statistique.*

Statistiques et estimateurs

- Une **statistique** est une valeur calculée à partir de l'échantillon (ex : moyenne, variance).
- Un **estimateur** est une formule utilisée pour **estimer un paramètre inconnu** de la population.

Exemple

On veut connaître la taille moyenne des étudiants d'une université.
$\rightarrow$ Population : tous les étudiants.
$\rightarrow$ Échantillon : 100 étudiants choisis au hasard.
$\rightarrow$ Estimateur : la moyenne de l'échantillon.

Estimation ponctuelle : cadre et notations

Objectif, hypothèses, exemples d'estimateurs

But : résumer une caractéristique **de population** par un **nombre unique** calculé sur l'échantillon.

Cadre : $X_1, \ldots, X_n$ i.i.d. de moyenne μ et variance σ^2 (échantillonnage simple).

Estimation ponctuelle : cadre et notations

Objectif, hypothèses, exemples d'estimateurs

But : résumer une caractéristique **de population** par un **nombre unique** calculé sur l'échantillon.

Cadre : $X_1, \ldots, X_n$ i.i.d. de moyenne μ et variance σ^2 (échantillonnage simple).

Exemples d'estimateurs ponctuels

- Moyenne : $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \rightsquigarrow \quad \mu$
- Variance (corrigée) : $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \rightsquigarrow \quad \sigma^2$
- Proportion : $\hat{p} = \frac{k}{n}$ où k = nb de succès $\quad \rightsquigarrow \quad P$

Estimation ponctuelle : cadre et notations

Objectif, hypothèses, exemples d'estimateurs

But : résumer une caractéristique **de population** par un **nombre unique** calculé sur l'échantillon.

Cadre : $X_1, \ldots, X_n$ i.i.d. de moyenne μ et variance σ^2 (échantillonnage simple).

Exemples d'estimateurs ponctuels

- Moyenne : $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \rightsquigarrow \quad \mu$
- Variance (corrigée) : $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \rightsquigarrow \quad \sigma^2$
- Proportion : $\hat{p} = \frac{k}{n}$ où k = nb de succès $\quad \rightsquigarrow \quad P$

Message clé

Une valeur ponctuelle est **utile mais incertaine** $\Rightarrow$ on mesurera son incertitude par la **variance** / **erreur standard**, puis par des **IC**.

Plan

1. Introduction à l'analyse inférencielle
2. Rappel mathématique : Loi normale, Khi-deux, Student, Fisher
3. Estimation par intervalle de confiance
4. Tests d'hypothèses

Introduction aux intervalles de confiance (IC)

- L'estimation ponctuelle fournit une valeur unique (ex : moyenne échantillonale $\bar{x}$), mais ne renseigne pas sur sa précision.
- Un **intervalle de confiance** (IC) donne une **plage plausible** pour le paramètre inconnu de la population.
- Exemple simple : moyenne des notes d'un échantillon de 30 élèves = 12, IC 95% = [11.5 ; 12.5].
- On interprète cela comme : « si on répète l'échantillonnage, 95% des IC contiendront la vraie valeur de la moyenne populationnelle. »

Idée intuitive des intervalles de confiance

- L'IC reflète l'incertitude liée à la taille de l'échantillon et à la dispersion des données.
- Plus l'échantillon est grand, plus l'IC est étroit.
- IC = $\bar{x} \pm$ marge d'erreur.
- Exemple graphique : plusieurs échantillons tirés $\rightarrow$ chaque échantillon donne un IC $\rightarrow$ 95% des IC contiennent la vraie moyenne.

Intervalle de confiance pour la moyenne (variance connue)

- Si la variance de la population σ^2 est connue :

$$IC_{1-\alpha} = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- Notation :
  - $\bar{x}$: moyenne de l'échantillon
  - n : taille de l'échantillon
  - $z_{\alpha/2}$: quantile de la loi normale
  - σ : écart-type population
- IC plus précis lorsque n est grand.

Intervalle de confiance pour la moyenne (variance inconnue)

- Quand σ est inconnue, utiliser l'écart-type échantillonnal s et la loi de Student :

$$IC_{1-\alpha} = \bar{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$$

- $t_{\alpha/2,n-1}$: quantile de Student à $n-1$ degrés de liberté.
- Plus adapté pour des petits échantillons ($n < 30$).

Exemple numérique : IC pour moyenne

Données : 30 élèves, $\bar{x} = 12$, $s = 1.5$, IC 95%

$$IC = \bar{x} \pm t_{0.025,29} \frac{s}{\sqrt{30}}$$

Calcul étape par étape :

1. $t_{0.025,29} \approx 2.045$
2. $s/\sqrt{n} = 1.5/\sqrt{30} \approx 0.274$
3. Marge d'erreur $= 2.045 * 0.274 \approx 0.56$

$$IC = 12 \pm 0.56 = [11.44; 12.56]$$

IC pour une proportion

- Si p est la proportion observée dans l'échantillon de taille n :

$$IC_{1-\alpha} = p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

- Exemple : 60% d'élèves réussissent sur $n = 50$:

$$IC_{95\%} = 0.6 \pm 1.96\sqrt{\frac{0.6 * 0.4}{50}} \approx [0.48; 0.72]$$

- Interprétation : on est 95% sûr que la proportion réelle se situe dans cet intervalle.

IC pour la variance / écart-type

- Pour la variance σ^2, avec n observations et s^2 échantillonale :

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}$$

- Exemple : $n = 10$, $s^2 = 4$, IC 95%
  - $\chi^2_{0.025,9} = 2.7$, $\chi^2_{0.975,9} = 19.0$
  - IC : $[9 * 4/19; 9 * 4/2.7] \approx [1.89; 13.33]$

Interprétation des IC

- L'IC exprime l'incertitude sur le paramètre réel.
- Plus l'échantillon est grand ou moins les données sont dispersées, plus l'IC est étroit.
- Comparer différents IC permet de visualiser la précision relative de différentes estimations.
- IC à 95% ne garantit pas que le paramètre est dans l'IC : il s'agit d'une probabilité sur la méthode d'estimation.

Résumé des formules des IC

Paramètre	**Formule IC**	**Remarques**
Moyenne, σ connue	$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$	Loi normale
Moyenne, σ inconnue	$\bar{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$	Loi de Student
Proportion	$p \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$	Intervalle dans [0,1]
Variance	$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}$	Loi du χ^2

Exemple complet avec interprétation

Échantillon : 20 élèves, moyenne $\bar{x} = 14$, écart-type $s = 2$, IC 95%.

$$t_{0.025,19} \approx 2.093, \quad \text{Erreur standard } = s/\sqrt{20} \approx 0.447$$

$$IC = 14 \pm 2.093 * 0.447 \approx [12.6; 15.4]$$

Interprétation :

- Si on prélève plusieurs échantillons de 20 élèves, 95% des IC construits contiendront la vraie moyenne de la population.
- L'intervalle [12.6 ; 15.4] indique que la moyenne réelle pourrait être légèrement plus faible ou plus élevée que la moyenne échantillonnale observée.

Plan

1. Introduction à l'analyse inférencielle
2. Rappel mathématique : Loi normale, Khi-deux, Student, Fisher
3. Estimation par intervalle de confiance
4. Tests d'hypothèses

Principes généraux des tests d'hypothèses

Définition générale

Un **test d'hypothèse** est une procédure statistique qui permet de prendre une décision concernant une population à partir d'un échantillon de données.

Hypothèse nulle et alternative

- **Hypothèse nulle (H_0)** : affirmation initiale que l'on cherche à tester (ex : la moyenne d'une population est égale à μ_0).
- **Hypothèse alternative (H_1)** : affirmation contraire à H_0 que l'on souhaite valider si H_0 est rejetée.

Erreurs dans les tests d'hypothèses

Erreur de type I

- Rejeter H_0 alors qu'elle est vraie.
- Probabilité de cette erreur : α, appelée **seuil de significativité** (souvent $\alpha = 5\%$).

Erreur de type II

- Ne pas rejeter H_0 alors qu'elle est fausse.
- Probabilité de cette erreur : β.
- Le pouvoir du test est $1 - \beta$, c'est la probabilité de détecter correctement une différence.

Seuil de significativité et p-value

Seuil de significativité α

- Fixé avant le test (ex : 0,05).
- Détermine la zone critique : si la statistique calculée tombe dans cette zone, H_0 est rejetée.

p-value

- Probabilité d'obtenir un résultat au moins aussi extrême que celui observé si H_0 est vraie.
- Interprétation :
  - $p \leq \alpha \Rightarrow$ rejeter H_0
  - $p > \alpha \Rightarrow$ ne pas rejeter H_0

Logique des tests d'hypothèses

Décision statistique

- On calcule une **statistique de test** à partir de l'échantillon.
- On compare cette statistique à la **valeur critique** correspondant à α.
- Si la statistique est trop improbable sous H_0, on rejette H_0 au profit de H_1.

Résumé

Condition	**Décision**
Statistique dans la zone critique	Rejeter H_0
Statistique dans la zone de non-rejet	Ne pas rejeter H_0

Tests sur la moyenne : Introduction

Objectif

Vérifier si la moyenne d'une population μ est égale à une valeur hypothétique μ_0 à partir d'un échantillon.

Cas principaux

- **Échantillon petit ($n < 30$) et variance inconnue :** Test de **Student (t)**.
- **Échantillon grand ($n \geq 30$) et variance connue :** Test **Z** (loi normale).

Hypothèses

Test bilatéral

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$$

- On teste si la moyenne diffère significativement de μ_0.

Test unilatéral

$$H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_1 : \mu > \mu_0$$

- On teste si la moyenne est significativement supérieure à μ_0.

Test sur la variance / écart-type : Introduction

Objectif

Vérifier si la variance d'une population σ^2 est égale à une valeur hypothétique σ_0^2 à partir d'un échantillon.

Idée générale

- Utiliser la statistique :

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

- Suit une loi du χ^2 avec $df = n - 1$ degrés de liberté.
- Permet de tester si les données sont plus ou moins dispersées que prévu.

Hypothèses et types de test

Test bilatéral

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{vs} \quad H_1 : \sigma^2 \neq \sigma_0^2$$

- On teste si la variance diffère de manière significative.

Test unilatéral

$$H_0 : \sigma^2 \leq \sigma_0^2 \quad \text{vs} \quad H_1 : \sigma^2 > \sigma_0^2$$

- On teste si la variance est significativement supérieure à σ_0^2.

Remarque

Le choix de la loi du χ^2 est adapté aux tests sur la variance car la statistique basée sur $(n-1)s^2/\sigma_0^2$ suit exactement cette loi si les données sont normalement distribuées.

Test d'égalité de deux moyennes : Introduction

Objectif

Comparer les moyennes de deux populations ou groupes pour déterminer si elles sont statistiquement différentes.

- Échantillons indépendants : les observations des deux groupes n'ont pas de lien.
- Échantillons appariés : chaque observation dans un groupe correspond à une observation dans l'autre groupe (ex : avant/après).
- La statistique de test suit une loi t (Student) ou t de Welch si variances inégales.

Hypothèses pour l'égalité de deux moyennes

Échantillons indépendants

- Bilatéral : $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$
- Unilatéral : $H_0 : \mu_1 \leq \mu_2$ vs $H_1 : \mu_1 > \mu_2$

Échantillons appariés

- On teste sur les différences $d_i = X_{1i} - X_{2i}$
- $H_0 : \bar{d} = 0$ vs $H_1 : \bar{d} \neq 0$

Test de Welch

Objectif

Comparer les moyennes de deux échantillons indépendants lorsque les variances des populations sont inégales. Le test de Student classique suppose l'égalité des variances. Si cette hypothèse est violée, le test de Welch est plus approprié.

- La statistique de test est similaire à celle de Student :

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Les degrés de liberté sont ajustés avec la formule de Welch :

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

- Approche robuste pour des échantillons de tailles différentes ou variances très différentes.

Tests de proportions : Introduction

Objectif

Les tests de proportions permettent de comparer des fréquences observées à des proportions théoriques ou entre deux groupes.

- Vérifier si une proportion observée diffère significativement d'une proportion théorique.
- Comparer deux proportions pour voir si elles sont statistiquement différentes.
- Approprié pour les variables qualitatives binaires (succès/échec, oui/non).

Test d'une proportion - Hypothèses

- Hypothèse nulle : $H_0 : p = p_0$
- Hypothèse alternative (bilatérale) : $H_1 : p \neq p_0$
- Statistique de test (approximation normale) :

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

où $\hat{p}$ est la proportion observée et n la taille de l'échantillon.

Comparaison de deux proportions

- Hypothèses :

$$H_0 : p_1 = p_2 \quad , \quad H_1 : p_1 \neq p_2$$

- Statistique de test :

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

où $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$ est la proportion combinée.

Interprétation

Points clés à retenir

- Les tests de proportions permettent de valider ou d'infirmer des hypothèses sur des fréquences observées.
- La statistique Z suit approximativement une loi normale pour un échantillon suffisamment grand.
- Les valeurs critiques et la p-value déterminent la décision statistique.
- Ces tests sont essentiels dans les évaluations pédagogiques, enquêtes ou études de marché.

ANOVA : Introduction

Objectif

L'ANOVA (Analysis of Variance) permet de comparer les moyennes de plusieurs groupes pour déterminer si elles diffèrent significativement.

- Alternative à effectuer plusieurs t : contrôle de l'erreur de type I.
- Principe : comparer la **variance entre les groupes** à la **variance intra-groupe**.
- Si la variance entre groupes est significativement plus grande que la variance intra-groupe, on rejette H_0.

ANOVA à un facteur

- Comparer k moyennes : $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$
- Hypothèses :

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_1 : \text{Au moins une moyenne diffère}$$

- Statistique de test :

$$F = \frac{\text{Variance entre groupes}}{\text{Variance intra-groupe}}$$

- Distribution sous H_0 : loi de Fisher $F_{k-1,N-k}$

ANOVA à deux facteurs

- Permet d'analyser l'effet de deux facteurs sur une variable quantitative et leur interaction.
- Exemples :
  - Sexe (H/F) et cycle d'étude (Licence/Master) sur la note finale.
- Hypothèses pour chaque facteur :

  H_0 : pas d'effet du facteur A, H_0 : pas d'effet du facteur B, H_0 : pas d'i

- Test basé sur variance entre groupes ajustée pour les autres facteurs.

Interprétation des résultats ANOVA

- Si $F_{obs} > F_{crit}$, on rejette H_0 pour l'effet testé.
- ANOVA à un facteur : si H_0 rejeté, au moins une classe diffère.
- ANOVA à deux facteurs : permet d'identifier si :
  - Un facteur a un effet significatif.
  - L'autre facteur a un effet significatif.
  - Il existe une interaction significative entre les deux facteurs.
- Les formules SSB, SSW, SSA, SSB, SSAB permettent de comprendre comment la variance est répartie.

Questions ?

Merci pour votre attention !