Estimation, Échantillonnage et Tests

H. Hocquard

HSE 2016-2017

Introduction : les 3 grandes lignes

Les statistiques peuvent permettre :

Introduction : les 3 grandes lignes

Les statistiques peuvent permettre :

- d'estimer un paramètre inconnu,

Introduction : les 3 grandes lignes

Les statistiques peuvent permettre :

- d'estimer un paramètre inconnu,
- de donner une zone dans laquelle un paramètre, a de grandes chances de se trouver,

Introduction : les 3 grandes lignes

Les statistiques peuvent permettre :

- d'estimer un paramètre inconnu,
- de donner une zone dans laquelle un paramètre, a de grandes chances de se trouver,
- de prendre des décisions.

Introduction : les 3 grandes lignes

Les statistiques peuvent permettre :

- d'estimer un paramètre inconnu,
- de donner une zone dans laquelle un paramètre, a de grandes chances de se trouver,
- de prendre des décisions.

Chacune de ses questions correspond à une thématique en statistiques.

Échantillonnage

L'échantillonnage permet de passer (de la loi connue d'un paramètre θ dans une population de taille N) à une estimée d'une quantité θ_n fabriquée à partir seulement d'une population de taille n plus petite (échantillon).

Échantillonnage

L'échantillonnage permet de passer (de la loi connue d'un paramètre θ dans une population de taille N) à une estimée d'une quantité θ_n fabriquée à partir seulement d'une population de taille n plus petite (échantillon).

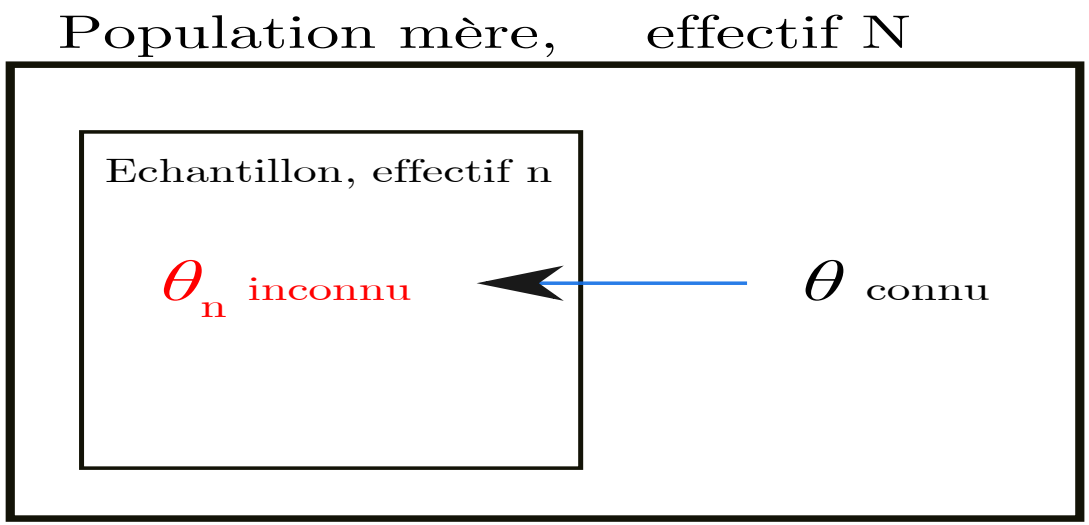

FIGURE: Principe de l'échantillonnage.

Exemple : échantillonnage

Exemple

Dans une entreprise qui compte 659 employés, on sait que $0,03\%$ des employés sont mécontents. On pioche un échantillon de 15 employés. Quel est l'ordre de grandeur des employés mécontents dans cet échantillon ?

Estimation

L'estimation permet d'induire, à partir des résulats observés sur un échantillon, des informations sur la population totale.

Estimation

L'estimation permet d'induire, à partir des résulats observés sur un échantillon, des informations sur la population totale.

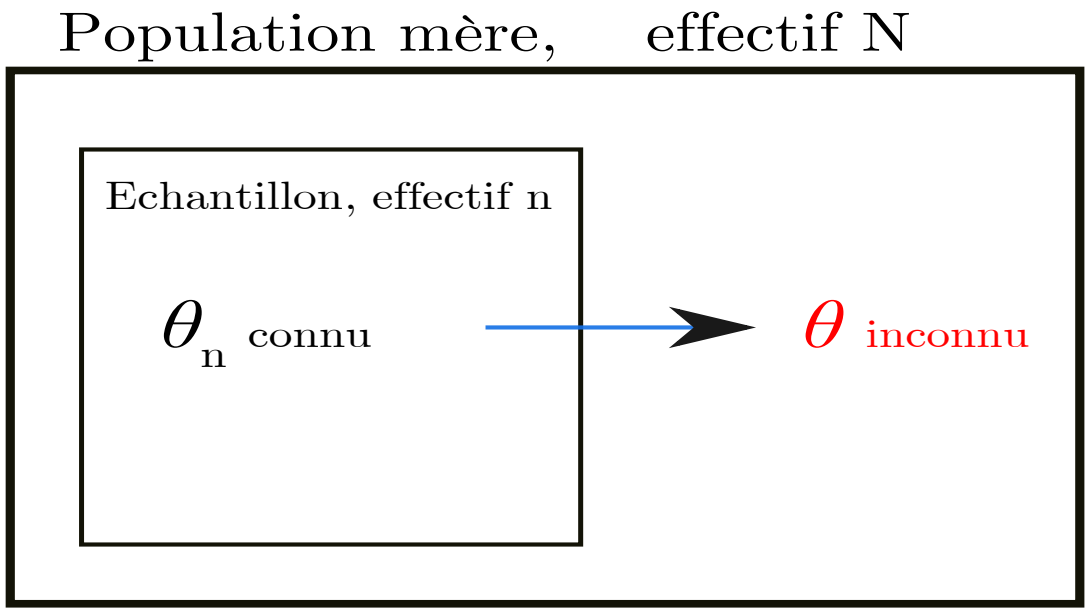

FIGURE: Principe de l'estimation.

Exemple : estimation

Exemple

Dans un échantillon de 15 employés d'une entreprise, 7% s'estiment sous pression. Quel est l'ordre de grandeur des employés sous pression parmi tout le personnel de l'entreprise ?

Tests statistiques

Tests de validité d'une hypothèse, prise de décision, contrôle qualité.

Tests statistiques

Tests de validité d'une hypothèse, prise de décision, contrôle qualité.

- Test sur un paramètre. Est-ce qu'une moyenne μ est inférieure à une valeur μ_0 ?

Tests statistiques

Tests de validité d'une hypothèse, prise de décision, contrôle qualité.

- Test sur un paramètre. Est-ce qu'une moyenne μ est inférieure à une valeur μ_0 ?
- Test de comparaison. Peut on considérer que la moyenne du chiffre d'affaire d'entreprises issues d'un réseau A, est la même que celle d'un réseau B ?

Tests statistiques

Tests de validité d'une hypothèse, prise de décision, contrôle qualité.

- Test sur un paramètre. Est-ce qu'une moyenne μ est inférieure à une valeur μ_0 ?
- Test de comparaison. Peut on considérer que la moyenne du chiffre d'affaire d'entreprises issues d'un réseau A, est la même que celle d'un réseau B ?

Tests statistiques

Tests de validité d'une hypothèse, prise de décision, contrôle qualité.

- Test sur un paramètre. Est-ce qu'une moyenne μ est inférieure à une valeur μ_0 ?
- Test de comparaison. Peut on considérer que la moyenne du chiffre d'affaire d'entreprises issues d'un réseau A, est la même que celle d'un réseau B ?

Étant donnée une marge d'erreur α, on rejettera ou ne rejettera pas une hypothèse au risque $\alpha\%$ de se tromper.

Tests statistiques

Tests de validité d'une hypothèse, prise de décision, contrôle qualité.

- Test sur un paramètre. Est-ce qu'une moyenne μ est inférieure à une valeur μ_0 ?
- Test de comparaison. Peut on considérer que la moyenne du chiffre d'affaire d'entreprises issues d'un réseau A, est la même que celle d'un réseau B ?

Étant donnée une marge d'erreur α, on rejettera ou ne rejettera pas une hypothèse au risque $\alpha\%$ de se tromper.

Remarque

On ne dira pas "qu'on valide une hypothèse" mais on dira "qu'on ne rejette pas une hypothèse". En effet, les théories probabilistes permettent de dire que sous une certaine hypothèse, il n'y a pas de contradictions...

Exemple : test d'un paramètre

- En vue d'aménager les heures de travail du personnel d'une entreprise, une étude s'est intéressée au temps de sommeil d'un échantillon des employés de l' entreprise.

Exemple : test d'un paramètre

- En vue d'aménager les heures de travail du personnel d'une entreprise, une étude s'est intéressée au temps de sommeil d'un échantillon des employés de l' entreprise.
- L'étude donne une moyenne du temps de sommeil de $6,56$ h et un écart-type de $1,35$ h.

Exemple : test d'un paramètre

- En vue d'aménager les heures de travail du personnel d'une entreprise, une étude s'est intéressée au temps de sommeil d'un échantillon des employés de l' entreprise.
- L'étude donne une moyenne du temps de sommeil de $6,56$ h et un écart-type de $1,35$ h.
- Peut-on considérer que le temps de sommeil des employés de cette entreprise est significativement inférieur au temps de sommeil moyen des individus qui est de 7h30 ?

Principe général commun

Dans chacun des cas, par les théorèmes probabilistes, on sait que :

- une quantité θ_n converge en loi vers une loi connue (loi normale, loi du χ^2, loi de Student, loi de Fisher, etc...en fonction des situations)

Principe général commun

Dans chacun des cas, par les théorèmes probabilistes, on sait que :

- une quantité θ_n converge en loi vers une loi connue (loi normale, loi du χ^2, loi de Student, loi de Fisher, etc...en fonction des situations)
- Par l'allure des densités de chacune de ces lois, on sait donc où la variable θ_n doit de trouver avec grosse probabilité...

Prérequis : lois classiques et convergence en loi

Lois limites classiques (que l'on obtiendra).
Connaître et savoir lire dans les tables les lois suivante :

- Loi normale $\mathcal{N}(0;1)$ et passage à $\mathcal{N}(\mu;\sigma)$,
- Loi du χ^2,
- Loi de Student,
- Loi de Fischer.

Rappel : convergence en loi.

On dit que la suite de v.a $(\theta_n)_n$ converge en loi vers la loi d'une v.a θ si,

pour tout intervalle $[a; b]$, on a :

$$\lim_{n \to +\infty} \mathbb{P}(\theta_n \in [a; b]) = \mathbb{P}(\theta \in [a; b]).$$

Rappel : convergence en loi.

On dit que la suite de v.a $(\theta_n)_n$ converge en loi vers la loi d'une v.a θ si,

pour tout intervalle $[a; b]$, on a :

$$\lim_{n \to +\infty} \mathbb{P}(\theta_n \in [a; b]) = \mathbb{P}(\theta \in [a; b]).$$

Notation : On écrit $\theta_n \xrightarrow{\mathcal{L}} \theta$

Rappel : convergence en loi.

On dit que la suite de v.a $(\theta_n)_n$ converge en loi vers la loi d'une v.a θ si,

pour tout intervalle $[a; b]$, on a :

$$\lim_{n \to +\infty} \mathbb{P}(\theta_n \in [a; b]) = \mathbb{P}(\theta \in [a; b]).$$

Notation : On écrit $\theta_n \xrightarrow{\mathcal{L}} \theta$

Exemple

Dans le Théorème central limite, on a vu que si les $(X_i)_i$ étaient iid et d'espérance finie μ et d'écart-type σ, alors la variable $\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)$ convergeait en loi vers une loi $\mathcal{N}(0; 1)$.

Décor, notations valables pour toute la suite du cours

Soit un échantillon de taille n.
Pour $1 \leq i \leq n$, notons X_i les valeurs d'un paramètre que prennent les n individus de l'échantillon.
Les X_i sont donc des v.a supposées i.i.d. (indépendantes et identiquement distribuées), de moyenne μ et d'écart-type σ (connus ou pas).

La moyenne empirique $\bar{X}_n$

On pose,

$$\bar{X}_n = \frac{\sum_{i=1\ldots n} X_i}{n}$$

Moyenne empirique des X_i.

Propriétés de $\bar{X}_n$

1. $$\mathbb{E}(\bar{X}_n) = \mu \quad et \quad Var(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Propriétés de $\bar{X}_n$

1. $$\mathbb{E}(\bar{X}_n) = \mu \quad et \quad Var(\bar{X}_n) = \frac{\sigma^2}{n}.$$

   (conséquences des propriétés de linéarité de l'espérance et de pseudo linéarité de la variance)

Propriétés de $\bar{X}_n$

1. $$\mathbb{E}(\bar{X}_n) = \mu \quad et \quad Var(\bar{X}_n) = \frac{\sigma^2}{n}.$$

   (conséquences des propriétés de linéarité de l'espérance et de pseudo linéarité de la variance)

2. $$\lim_n \bar{X}_n = \mu$$

   (Loi des grands nombres)

Propriétés de $\bar{X}_n$

1. $$\mathbb{E}(\bar{X}_n) = \mu \quad et \quad Var(\bar{X}_n) = \frac{\sigma^2}{n}.$$

   (conséquences des propriétés de linéarité de l'espérance et de pseudo linéarité de la variance)

2. $$\lim_n \bar{X}_n = \mu$$

   (Loi des grands nombres)

3. $$\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0;1).$$

   (Théorème central limite)

Justification de l'intérêt de la quantité $\bar{X}_n$

- Ainsi, la statistique $\bar{X}_n$ converge (en un certain sens) quand n tend vers l'infini vers $\mu = \mathbb{E}(X_i)$. On dit que c'est un estimateur de μ.

Justification de l'intérêt de la quantité $\bar{X}_n$

- Ainsi, la statistique $\bar{X}_n$ converge (en un certain sens) quand n tend vers l'infini vers $\mu = \mathbb{E}(X_i)$. On dit que c'est un estimateur de μ.
- On dit qu'il est sans biais car, $\mathbb{E}(\bar{X}_n) = \mu$ (l'espérance de l'estimateur est égale à la valeur que l'on cherche à estimer).

Exemple d'application

Exemple

Parmi le personnel d'une entreprise, il y a 300 femmes et 600 hommes. On réalise une enquête sur un échantillon de 55 personnes. Donnez une fourchette du nombres d'hommes et de femmes de l'échantillon, avec proba 0,95.

Variance empirique

Au regard de la définition de la variance et de la loi des grands nombres, il est naturel d'introduire :

$$S_n^2 = \frac{1}{n} \sum_{i=1\ldots n} (X_i - \bar{X}_n)^2 = \frac{1}{n}[\sum_{i=1\ldots n} X_i^2\ ] - \bar{X}_n{}^2$$

la variance empirique des X_i.

Variance empirique

Au regard de la définition de la variance et de la loi des grands nombres, il est naturel d'introduire :

$$S_n^2 = \frac{1}{n} \sum_{i=1\ldots n} (X_i - \bar{X}_n)^2 = \frac{1}{n}[\sum_{i=1\ldots n} X_i^2\ ] - \bar{X}_n{}^2$$

la variance empirique des X_i.

Avantage de cet estimateur, on n'a pas besoin de connaître l'espérance μ.

Propriétés

1. $$\lim_n S_n^2 = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 = \sigma^2$$

Propriétés

1. $$\lim_n S_n^2 = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 = \sigma^2$$

   (loi des grands nombres aux X_i^2 et X_i)

Propriétés

1. $$\lim_n S_n^2 = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 = \sigma^2$$

   (loi des grands nombres aux X_i^2 et X_i)

2. $$\mathbb{E}(S_n^2) = \frac{n-1}{n}\sigma^2$$

Propriétés

1. $$\lim_n S_n^2 = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 = \sigma^2$$

   (loi des grands nombres aux X_i^2 et X_i)

2. $$\mathbb{E}(S_n^2) = \frac{n-1}{n}\sigma^2$$

   (petit calcul) On dit que S_n^2 a un biais, $\mathbb{E}(S_n^2) \neq \sigma^2$.

3. On admet que,

$$\frac{S_n^2 - \frac{n-1}{n}\sigma^2}{Var(S_n^2)} \xrightarrow{\mathcal{L}} \mathcal{N}(0;1).$$

Variance empirique corrigée

Au regard de la définition de la variance et de la loi des grands nombres, il est naturel d'introduire :

$$\hat{S}^2_{n-1} = \frac{1}{n-1} \sum_{i=1\ldots n} (X_i - \bar{X}_n)^2$$

l'estimateur sans biais de la variance.

Variance empirique corrigée

Au regard de la définition de la variance et de la loi des grands nombres, il est naturel d'introduire :

$$\hat{S}^2_{n-1} = \frac{1}{n-1} \sum_{i=1\ldots n} (X_i - \bar{X}_n)^2$$

l'estimateur sans biais de la variance.

Avantage de cet estimateur : $\hat{S}^2_{n-1}$ est sans biais, puisque $\mathbb{E}(\hat{S}^2_{n-1}) = \sigma^2$.

Outil : estimateur

- Pour estimer une quantité, on fabrique une fonction des observations, (une statistique/une v.a) qui tend vers la quantité souhaitée, à l'aide des théorémes limites (type Loi des grands nombres). On préfèrera une statistique sans biais.

Outil : estimateur

- Pour estimer une quantité, on fabrique une fonction des observations, (une statistique/une v.a) qui tend vers la quantité souhaitée, à l'aide des théorémes limites (type Loi des grands nombres). On préfèrera une statistique sans biais.
- On essaie de connaître la loi limite de cette statistique.

Outil : estimateur

- Pour estimer une quantité, on fabrique une fonction des observations, (une statistique/une v.a) qui tend vers la quantité souhaitée, à l'aide des théorémes limites (type Loi des grands nombres). On préfèrera une statistique sans biais.
- On essaie de connaître la loi limite de cette statistique.
- On est alors capable, de donner les fluctuations les plus probables de la statistique, et de donner par exemple un intervalle de confiance.

Estimation d'une moyenne, lorsque σ connu

Estimation d'une moyenne, dans le cas où la variance σ^2 est connue

- On estime la moyenne μ par $\bar{X}_n$.

Estimation d'une moyenne, dans le cas où la variance σ^2 est connue

- On estime la moyenne μ par $\bar{X}_n$.
- Par le TCL, on sait que $\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0;1)$.

Estimation d'une moyenne, dans le cas où la variance σ^2 est connue

- On estime la moyenne μ par $\bar{X}_n$.
- Par le TCL, on sait que $\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0;1)$.
- Étant donnée une marge d'erreur α, (par ex 5%), on détermine alors un certain u_α à l'aide de la table de la $\mathcal{N}(0;1)$, tel que

$$\mathbb{P}(|Z| \leq u_\alpha) \geq 1 - \alpha, \qquad \text{où} \quad Z \sim \mathcal{N}(0;1).$$

Estimation d'une moyenne, dans le cas où la variance σ^2 est connue

- On estime la moyenne μ par $\bar{X}_n$.
- Par le TCL, on sait que $\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0;1)$.
- Étant donnée une marge d'erreur α, (par ex 5%), on détermine alors un certain u_α à l'aide de la table de la $\mathcal{N}(0;1)$, tel que

$$\mathbb{P}(|Z| \leq u_\alpha) \geq 1 - \alpha, \qquad \text{où} \quad Z \sim \mathcal{N}(0;1).$$

  On pourra remarquer que $u_\alpha = z_{1-\frac{\alpha}{2}} = \Pi^{-1}(1 - \frac{\alpha}{2})$.
- Ainsi avec proba $\geq 1 - \alpha$, on a :$|\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)| \leq z_{1-\frac{\alpha}{2}}$,

Estimation d'une moyenne, dans le cas où la variance σ^2 est connue

- On estime la moyenne μ par $\bar{X}_n$.
- Par le TCL, on sait que $\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0;1)$.
- Étant donnée une marge d'erreur α, (par ex 5%), on détermine alors un certain u_α à l'aide de la table de la $\mathcal{N}(0;1)$, tel que

$$\mathbb{P}(|Z| \leq u_\alpha) \geq 1 - \alpha, \qquad \text{où} \quad Z \sim \mathcal{N}(0;1).$$

  On pourra remarquer que $u_\alpha = z_{1-\frac{\alpha}{2}} = \Pi^{-1}(1 - \frac{\alpha}{2})$.
- Ainsi avec proba $\geq 1 - \alpha$, on a : $|\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)| \leq z_{1-\frac{\alpha}{2}}$, *i.e.* :

$$\bar{X}_n - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Exemple

Exemple

Une machine produit en grande série des objets de masse théorique 180g. On admet que la variable aléatoire qui associe à un objet sa masse a pour écart-type 0,92g. On prélève un échantillon de 100 objets et on mesure la masse de chacun, on obtient une moyenne de 179,93g. Déterminer un intervalle de confiance au seuil de risque de 1%, de la masse μ d'un objet.

Exemple

- Soit X_i, la v.a qui renvoit la masse de l'objet i de l'échantillon. On cherche un intervalle de confiance de $\mu = \mathbb{E}(X_i)$.

Exemple

- Soit X_i, la v.a qui renvoit la masse de l'objet i de l'échantillon. On cherche un intervalle de confiance de $\mu = \mathbb{E}(X_i)$.
- On sait qu'avec proba $1 - \alpha$,

$$\bar{X_n} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X_n} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

- $\alpha = 0,01$ donne un $z_{1-\frac{\alpha}{2}} = 2,58$ (table de la loi normale centrée réduite).

Exemple

- Soit X_i, la v.a qui renvoit la masse de l'objet i de l'échantillon. On cherche un intervalle de confiance de $\mu = \mathbb{E}(X_i)$.
- On sait qu'avec proba $1 - \alpha$,

$$\bar{X_n} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X_n} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

- $\alpha = 0,01$ donne un $z_{1-\frac{\alpha}{2}} = 2,58$ (table de la loi normale centrée réduite).
- D'où,

$$179,93 - 2,58 \times \frac{0,92}{\sqrt{100}} \leq \mu \leq 179,93 + 2,58 \times \frac{0,92}{\sqrt{100}}$$

i.e. :

Avec proba $0,99$ on a, $\mu \in [179,69; 180,17]$.

Estimation d'une moyenne, lorsque σ inconnu

Estimation de la moyenne, dans le cas où la variance σ^2 est inconnue.

On suppose dans ce cas que les X_i suivent des lois normales $\mathcal{N}(\mu, \sigma)$.

Estimation de la moyenne, dans le cas où la variance σ^2 est inconnue.

- On utilise le fait que $T = \frac{\sqrt{n}}{\hat{S}_{n-1}}(\bar{X}_n - \mu)$ suit une loi de $Student(n-1)$.

Estimation de la moyenne, dans le cas où la variance σ^2 est inconnue.

- On utilise le fait que $T = \frac{\sqrt{n}}{\hat{S}_{n-1}}(\bar{X}_n - \mu)$ suit une loi de $Student(n-1)$.
- Étant donnée une marge d'erreur α, on détermine alors un certain t_α à l'aide de la table de la loi de student, tel que

$$\mathbb{P}(|T| \leq t_\alpha) \geq 1 - \alpha, \quad \text{où} \quad T \sim Student(n-1).$$

Estimation de la moyenne, dans le cas où la variance σ^2 est inconnue.

- On utilise le fait que $T = \frac{\sqrt{n}}{\hat{S}_{n-1}}(\bar{X}_n - \mu)$ suit une loi de $Student(n-1)$.
- Étant donnée une marge d'erreur α, on détermine alors un certain t_α à l'aide de la table de la loi de student, tel que

$$\mathbb{P}(|T| \leq t_\alpha) \geq 1 - \alpha, \quad \text{où} \quad T \sim Student(n-1).$$

- On conclut, qu'avec proba au moins $1-\alpha$, on a :

$$\left| \frac{\sqrt{n}}{\hat{S}_{n-1}}(\bar{X}_n - \mu) \right| \leq t_\alpha.$$

Ainsi,

Estimation de la moyenne, dans le cas où la variance σ^2 est inconnue.

- On utilise le fait que $T = \frac{\sqrt{n}}{\hat{S}_{n-1}}(\bar{X}_n - \mu)$ suit une loi de $Student(n-1)$.
- Étant donnée une marge d'erreur α, on détermine alors un certain t_α à l'aide de la table de la loi de student, tel que

$$\mathbb{P}(|T| \leq t_\alpha) \geq 1 - \alpha, \quad \text{où} \quad T \sim Student(n-1).$$

- On conclut, qu'avec proba au moins $1-\alpha$, on a :

$$\left| \frac{\sqrt{n}}{\hat{S}_{n-1}}(\bar{X}_n - \mu) \right| \leq t_\alpha.$$

Ainsi,

$$\bar{X}_n - t_\alpha \frac{\hat{S}_{n-1}}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_\alpha \frac{\hat{S}_{n-1}}{\sqrt{n}}$$

Exemple

Exemple

Le chiffre d'affaire mensuel d'une entreprise suit une loi normale de moyenne μ et d'écart-type σ inconnus. Sur les 12 derniers mois, on a observé une moyenne des chiffres d'affaires égale à 10 000 euros avec un écart-type de 2000 euros. Donner une estimation de μ par intervalle de confiance au niveau $0,98$.

Exemple

- Soit X_i le chiffre d'affaire de l'entreprise le mois i.

Exemple

- Soit X_i le chiffre d'affaire de l'entreprise le mois i.
- On sait que $T = \frac{\sqrt{11}}{S_{12}}(\bar{X_{12}} - \mu)$ suit une loi de $Student(11)$.

Exemple

- Soit X_i le chiffre d'affaire de l'entreprise le mois i.
- On sait que $T = \frac{\sqrt{11}}{S_{12}}(\bar{X_{12}} - \mu)$ suit une loi de $Student(11)$.
- À l'aide de la table de la loi de Student, on trouve $t_\alpha = t_{0,02} \simeq 2,718$ tel que $\mathbb{P}(|T| \leq 2,718) \geq 0,98$.

Exemple

- Soit X_i le chiffre d'affaire de l'entreprise le mois i.
- On sait que $T = \frac{\sqrt{11}}{S_{12}}(\bar{X_{12}} - \mu)$ suit une loi de $Student(11)$.
- À l'aide de la table de la loi de Student, on trouve $t_\alpha = t_{0,02} \simeq 2,718$ tel que $\mathbb{P}(|T| \leq 2,718) \geq 0,98$.
- Donc, $|\frac{\sqrt{11}}{S_{12}}(\bar{X_{12}} - \mu)| \leq 2,718$ avec proba $\geq 0,98$.

Exemple

- Soit X_i le chiffre d'affaire de l'entreprise le mois i.
- On sait que $T = \frac{\sqrt{11}}{S_{12}}(\bar{X_{12}} - \mu)$ suit une loi de $Student(11)$.
- À l'aide de la table de la loi de Student, on trouve $t_\alpha = t_{0,02} \simeq 2,718$ tel que $\mathbb{P}(|T| \leq 2,718) \geq 0,98$.
- Donc, $|\frac{\sqrt{11}}{S_{12}}(\bar{X_{12}} - \mu)| \leq 2,718$ avec proba $\geq 0,98$.

$$i.e. : \quad \mu \in [\bar{X_{12}} - 2,718 \times \frac{S_{12}}{\sqrt{11}}; \bar{X_{12}} + 2,718 \times \frac{S_{12}}{\sqrt{11}}].$$

Exemple

- Soit X_i le chiffre d'affaire de l'entreprise le mois i.
- On sait que $T = \frac{\sqrt{11}}{S_{12}}(\bar{X_{12}} - \mu)$ suit une loi de $Student(11)$.
- À l'aide de la table de la loi de Student, on trouve $t_\alpha = t_{0,02} \simeq 2,718$ tel que $\mathbb{P}(|T| \leq 2,718) \geq 0,98$.
- Donc, $|\frac{\sqrt{11}}{S_{12}}(\bar{X_{12}} - \mu)| \leq 2,718$ avec proba $\geq 0,98$.

$$i.e. : \quad \mu \in [\bar{X_{12}} - 2,718 \times \frac{S_{12}}{\sqrt{11}}; \bar{X_{12}} + 2,718 \times \frac{S_{12}}{\sqrt{11}}].$$

Avec $\bar{X_{12}} = 10000$ et $S_{12} = 2000$, on obtient

$$\mu \in [8360,9; 11639,02], \quad \text{avec proba } 0,98.$$

Intervalle de confiance d'une proportion

Intervalle de confiance d'une proportion

Soit A un évènement aléatoire de proba p.
On appelle $\hat{p} = \frac{1}{n} \times$ nombre de fois où X_i réalise A.
$\hat{p}$ est un estimateur sans biais de p.
On peut alors en déduire :

$$\hat{p} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Exemple

Exemple

Dans un échantillon de 197 pommes, on constate que 19 d'entre elles sont abimées. Déterminer un intervalle de confiance au risque 5% de la proportion de pommes abimées dans la production.

Exemple

- On sait qu'avec proba $1-\alpha$,

$$\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- $\alpha = 0,05$ donne un $z_{1-\frac{\alpha}{2}} = 1,96$ (table de la loi normale centrée réduite).

Exemple

- On sait qu'avec proba $1-\alpha$,

$$\hat{p} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- $\alpha = 0,05$ donne un $z_{1-\frac{\alpha}{2}} = 1,96$ (table de la loi normale centrée réduite).
- D'où,

$$\frac{19}{197} - 1,96 \times 0,021 \leq p \leq \frac{19}{197} + 1,96 \times 0,021$$

i.e. :

Avec proba $0,95$ on a, $\quad p \in [0,055; 0,137]$.

Tests paramétriques

Point de vue global

- C'est une stratégie analogue à celles des estimations. On utilise la même technologie.

Point de vue global

- C'est une stratégie analogue à celles des estimations. On utilise la même technologie.
- On fait une hypothèse sur un paramètre. On sait alors que, sous cette hypothèse, une certaine v.a (une statistique) doit être distribuée suivant une certaine loi (Théorème limites, distribution d'échantillonage).

Point de vue global

- C'est une stratégie analogue à celles des estimations. On utilise la même technologie.
- On fait une hypothèse sur un paramètre. On sait alors que, sous cette hypothèse, une certaine v.a (une statistique) doit être distribuée suivant une certaine loi (Théorème limites, distribution d'échantillonage).
- On vérifie alors, avec un taux α, "l'adéquation" des 2 lois. Il existe des tests unilatéraux et bilatéraux.

Test bilatéral d'une moyenne, σ connu

Test bilatéral d'une moyenne, cas où σ connu

- On veut tester l'hypothèse $H_0 : \ \mu = \mu_0$ contre $H_1 : \ \mu \neq \mu_0$.

Test bilatéral d'une moyenne, cas où σ connu

- On veut tester l'hypothèse $H_0 : \ \mu = \mu_0$ contre $H_1 : \ \mu \neq \mu_0$.
- Sous H_0, on sait que pour n grand, la v.a

$$\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu_0) \text{ doit suivre une } \mathcal{N}(0;1).$$

Test bilatéral d'une moyenne, cas où σ connu

- On veut tester l'hypothèse $H_0 : \ \mu = \mu_0$ contre $H_1 : \ \mu \neq \mu_0$.
- Sous H_0, on sait que pour n grand, la v.a

$$\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu_0) \text{ doit suivre une } \mathcal{N}(0;1).$$

- Étant donnée une marge d'erreur α, on détermine alors un certain u_α à l'aide de la table de la $\mathcal{N}(0;1)$, tel que

$$\mathbb{P}(|Z| \leq u_\alpha) \geq 1 - \alpha, \qquad \text{où} \quad Z \sim \mathcal{N}(0;1).$$

Test bilatéral d'une moyenne, cas où σ connu

- On veut tester l'hypothèse $H_0 : \ \mu = \mu_0$ contre $H_1 : \ \mu \neq \mu_0$.
- Sous H_0, on sait que pour n grand, la v.a

$$\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu_0) \text{ doit suivre une } \mathcal{N}(0;1).$$

- Étant donnée une marge d'erreur α, on détermine alors un certain u_α à l'aide de la table de la $\mathcal{N}(0;1)$, tel que

$$\mathbb{P}(|Z| \leq u_\alpha) \geq 1 - \alpha, \qquad \text{où} \ \ Z \sim \mathcal{N}(0;1).$$

  En fait, $u_\alpha = z_{1-\frac{\alpha}{2}}$.
- La position du nombre $\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu_0)$ par rapport à $[-z_{1-\frac{\alpha}{2}}; z_{1-\frac{\alpha}{2}}]$, permet de rejeter ou de ne pas rejeter H_0.

Test bilatéral d'une moyenne, cas où σ connu

Ainsi,

- Si $\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu_0) \in [-z_{1-\frac{\alpha}{2}}; z_{1-\frac{\alpha}{2}}]$, on ne rejette pas H_0.

Test bilatéral d'une moyenne, cas où σ connu

Ainsi,

- Si $\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu_0) \in [-z_{1-\frac{\alpha}{2}}; z_{1-\frac{\alpha}{2}}]$, on ne rejette pas H_0.
- Sinon on rejette H_0.

Test bilatéral d'une moyenne, cas où σ connu

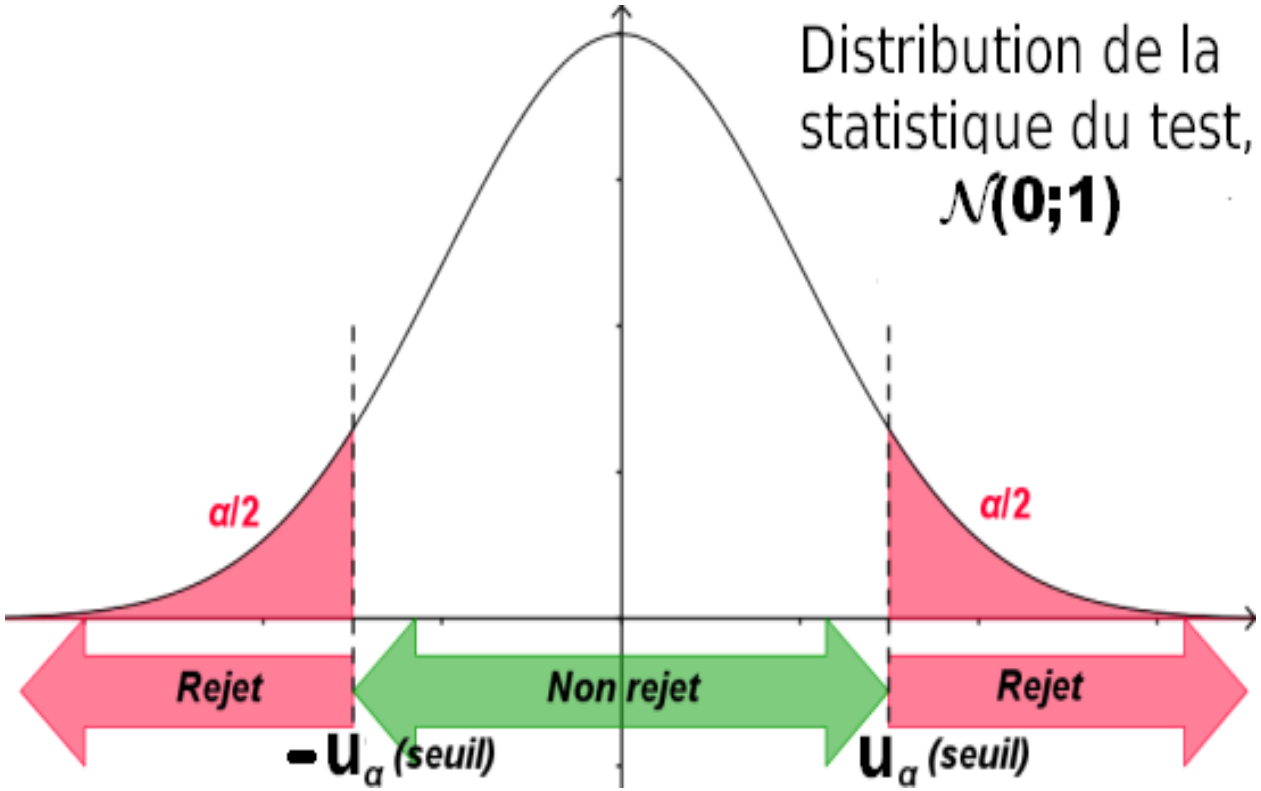

FIGURE: Zones de rejet et de non rejet de H_0.

Test bilatéral d'une moyenne, σ inconnu mais échantillon Gaussien

Test bilatéral d'une moyenne, cas d'un échantillon Gaussien et σ inconnu

Dans le cas où σ est inconnu, c'est le même principe mais on raisonne cette fois avec la variable de décision T_{n-1} et on suppose que les X_i suivent une loi Normale $\mathcal{N}(\mu; \sigma)$.

Test bilatéral d'une moyenne, cas d'un échantillon Gaussien et σ inconnu

- On veut tester l'hypothèse $H_0 : \ \mu = \mu_0$ contre $H_1 : \ \mu \neq \mu_0$.

Test bilatéral d'une moyenne, cas d'un échantillon Gaussien et σ inconnu

- On veut tester l'hypothèse $H_0 : \ \mu = \mu_0$ contre $H_1 : \ \mu \neq \mu_0$.
- Sous H_0, on sait que la v.a,

$$T_{n-1} = \frac{\sqrt{n}}{\hat{S}_{n-1}}(\bar{X}_n - \mu_0) \text{ doit suit une loi de } \mathit{Student}(n-1)$$

Test bilatéral d'une moyenne, cas d'un échantillon Gaussien et σ inconnu

- On veut tester l'hypothèse $H_0 : \ \mu = \mu_0$ contre $H_1 : \ \mu \neq \mu_0$.
- Sous H_0, on sait que la v.a,

$$T_{n-1} = \frac{\sqrt{n}}{\hat{S}_{n-1}}(\bar{X}_n - \mu_0) \text{ doit suit une loi de } \mathit{Student}(n-1)$$

- Étant donnée une marge d'erreur α, on détermine alors un certain t_α à l'aide de la table de $\mathit{Student}(n-1)$, tel que

$$\mathbb{P}(|T_{n-1}| \leq t_\alpha) \geq 1 - \alpha.$$

Test bilatéral d'une moyenne, cas d'un échantillon Gaussien et σ inconnu

- On veut tester l'hypothèse $H_0 : \ \mu = \mu_0$ contre $H_1 : \ \mu \neq \mu_0$.
- Sous H_0, on sait que la v.a,

$$T_{n-1} = \frac{\sqrt{n}}{\hat{S}_{n-1}}(\bar{X}_n - \mu_0) \text{ doit suit une loi de } \mathit{Student}(n-1)$$

- Étant donnée une marge d'erreur α, on détermine alors un certain t_α à l'aide de la table de $\mathit{Student}(n-1)$, tel que

$$\mathbb{P}(|T_{n-1}| \leq t_\alpha) \geq 1 - \alpha.$$

- Étude de la position de $\frac{\sqrt{n}}{\hat{S}_{n-1}}(\bar{X}_n - \mu_0)$ par rapport à $[-t_\alpha; t_\alpha]$.

Test bilatéral d'une moyenne, cas d'un échantillon Gaussien et σ inconnu

Ainsi, on a la même discussion,

- Si $T_{n-1} \in [-t_\alpha; t_\alpha]$, on ne rejette pas H_0.

Test bilatéral d'une moyenne, cas d'un échantillon Gaussien et σ inconnu

Ainsi, on a la même discussion,

- Si $T_{n-1} \in [-t_\alpha; t_\alpha]$, on ne rejette pas H_0.
- Sinon on rejette H_0.

Test bilatéral d'une moyenne, cas d'un échantillon Gaussien et σ inconnu

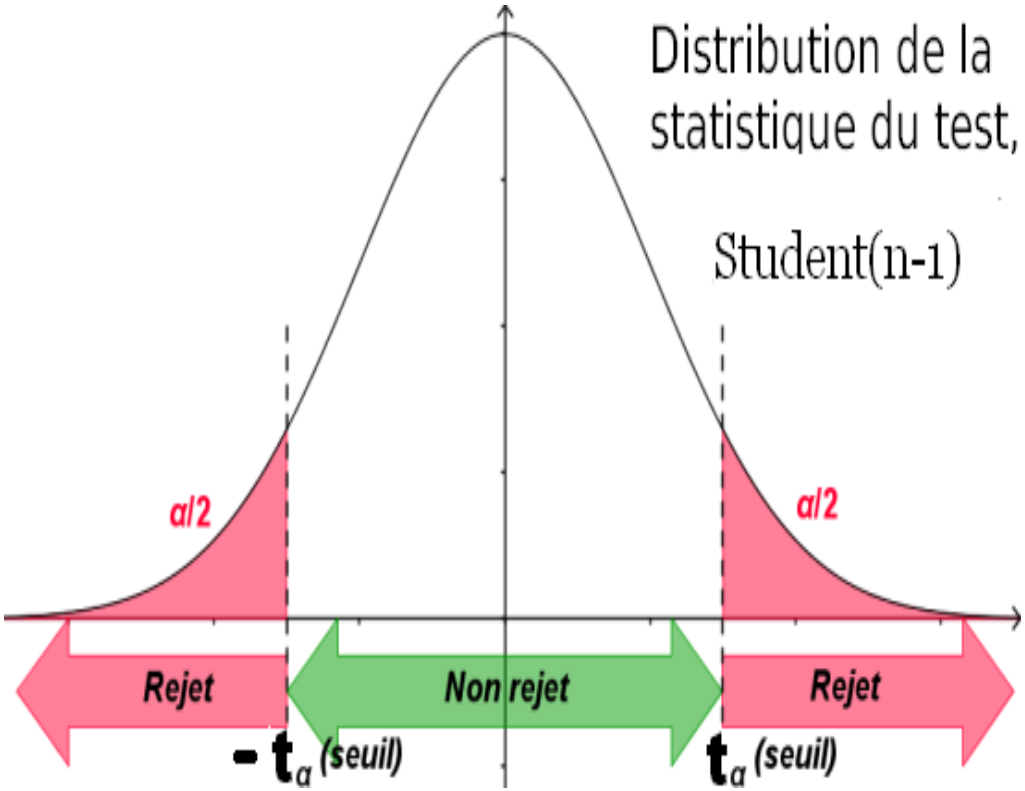

FIGURE: Zones de rejet et de non rejet de H_0.

Exemple : Effet sur le temps de sommeil des aménagements d'horaires dans une entreprise

- En vue d'aménager les heures de travail du personnel d'une entreprise, une étude s'est interessée au temps de sommeil d'un échantillon de 30 employés de l'entreprise.

Exemple : Effet sur le temps de sommeil des aménagements d'horaires dans une entreprise

- En vue d'aménager les heures de travail du personnel d'une entreprise, une étude s'est interessée au temps de sommeil d'un échantillon de 30 employés de l'entreprise.
- L'étude donne une moyenne du temps de sommeil de $6,56$ h et un écart-type de $1,35$ h.

Exemple : Effet sur le temps de sommeil des aménagements d'horaires dans une entreprise

- En vue d'aménager les heures de travail du personnel d'une entreprise, une étude s'est interessée au temps de sommeil d'un échantillon de 30 employés de l'entreprise.
- L'étude donne une moyenne du temps de sommeil de $6,56$ h et un écart-type de $1,35$ h.
- En supposant que le temps de sommeil d'un employé suit une loi normale, peut on considérer que le temps de sommeil des employés de cette entreprise est inférieur au temps de sommeil moyen des individus qui est de 7h30 au seuil 5% ?

Exemple : Effet sur le temps de sommeil des aménagements d' horaire dans une entreprise

- Soit X_i le temps de sommeil de la personne i de l'échantillon, on suppose que $X_i \sim \mathcal{N}(\mu, \sigma)$.

Exemple : Effet sur le temps de sommeil des aménagements d' horaire dans une entreprise

- Soit X_i le temps de sommeil de la personne i de l'échantillon, on suppose que $X_i \sim \mathcal{N}(\mu, \sigma)$.
- Soit $H_0 : \mu = 7,5$.

Exemple : Effet sur le temps de sommeil des aménagements d' horaire dans une entreprise

- Soit X_i le temps de sommeil de la personne i de l'échantillon, on suppose que $X_i \sim \mathcal{N}(\mu, \sigma)$.
- Soit $H_0 : \mu = 7,5$.
- Sous H_0, on a $T_{29} = \frac{\sqrt{29}}{S_{30}}(\bar{X_{30}} - 7,5) \sim Student(29)$.

Exemple : Effet sur le temps de sommeil des aménagements d' horaire dans une entreprise

- Soit X_i le temps de sommeil de la personne i de l'échantillon, on suppose que $X_i \sim \mathcal{N}(\mu, \sigma)$.
- Soit $H_0 : \mu = 7,5$.
- Sous H_0, on a $T_{29} = \frac{\sqrt{29}}{S_{30}}(\bar{X_{30}} - 7,5) \sim Student(29)$.
- La table de Student donne $\mathbb{P}(|T| \leq 2,045) \geq 0,95$.

Exemple : Effet sur le temps de sommeil des aménagements d' horaire dans une entreprise

- Soit X_i le temps de sommeil de la personne i de l'échantillon, on suppose que $X_i \sim \mathcal{N}(\mu, \sigma)$.
- Soit $H_0 : \mu = 7,5$.
- Sous H_0, on a $T_{29} = \frac{\sqrt{29}}{S_{30}}(\bar{X_{30}} - 7,5) \sim Student(29)$.
- La table de Student donne $\mathbb{P}(|T| \leq 2,045) \geq 0,95$.
- Ici $T_{29} = \frac{\sqrt{29}}{1,35}(6,56 - 7,5) = -3,74$.

Exemple : Effet sur le temps de sommeil des aménagements d' horaire dans une entreprise

- Soit X_i le temps de sommeil de la personne i de l'échantillon, on suppose que $X_i \sim \mathcal{N}(\mu, \sigma)$.
- Soit $H_0 : \mu = 7,5$.
- Sous H_0, on a $T_{29} = \frac{\sqrt{29}}{S_{30}}(\bar{X_{30}} - 7,5) \sim Student(29)$.
- La table de Student donne $\mathbb{P}(|T| \leq 2,045) \geq 0,95$.
- Ici $T_{29} = \frac{\sqrt{29}}{1,35}(6,56 - 7,5) = -3,74$.
- Donc $T_{29} \notin [-2,045 : 2,045]$, et on rejette H_0.

Test unilatéral d'une moyenne, cas d'un échantillon Gaussien et σ connu

Test unilatéral d'une moyenne, cas d'un échantillon Gaussien et σ connu

Ainsi, on a la même discussion,

- On veut tester l'hypothèse $H_0 : \ \mu < \mu_0$ contre $H_1 : \ \mu \geq \mu_0$.

Test unilatéral d'une moyenne, cas d'un échantillon Gaussien et σ connu

Ainsi, on a la même discussion,

- On veut tester l'hypothèse $H_0 : \ \mu < \mu_0$ contre $H_1 : \ \mu \geq \mu_0$.
- Si $\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu_0) \in [-z_{1-\frac{\alpha}{2}}; z_{1-\frac{\alpha}{2}}]$, on ne rejette pas H_0.

Test unilatéral d'une moyenne, cas d'un échantillon Gaussien et σ connu

Ainsi, on a la même discussion,

- On veut tester l'hypothèse $H_0 : \ \mu < \mu_0$ contre $H_1 : \ \mu \geq \mu_0$.
- Si $\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu_0) \in [-z_{1-\frac{\alpha}{2}}; z_{1-\frac{\alpha}{2}}]$, on ne rejette pas H_0.
- Sinon on rejette H_0.

Test unilatéral d'une moyenne, cas d'un échantillon Gaussien et σ inconnu

Test unilatéral d'une moyenne, cas d'un échantillon Gaussien et σ inconnu

- On veut tester l'hypothèse $H_0 : \ \mu < \mu_0$ contre $H_1 : \ \mu \geq \mu_0$.

Test unilatéral d'une moyenne, cas d'un échantillon Gaussien et σ inconnu

- On veut tester l'hypothèse $H_0 : \ \mu < \mu_0$ contre $H_1 : \ \mu \geq \mu_0$.
- Si $T_{n-1} = \frac{\sqrt{n}}{\hat{S}_{n-1}}(\bar{X}_n - \mu_0) \in [-t_{2\alpha}; t_{2\alpha}]$, on ne rejette pas H_0.

Test unilatéral d'une moyenne, cas d'un échantillon Gaussien et σ inconnu

- On veut tester l'hypothèse $H_0 : \ \mu < \mu_0$ contre $H_1 : \ \mu \geq \mu_0$.
- Si $T_{n-1} = \frac{\sqrt{n}}{\hat{S}_{n-1}}(\bar{X}_n - \mu_0) \in [-t_{2\alpha}; t_{2\alpha}]$, on ne rejette pas H_0.
- Sinon on rejette H_0.

Test d'une proportion

Test d'une proportion

 On suppose dans ce cas que les X_i suivent des lois de Bernouilli $\mathcal{B}(p)$. On souhaite comparer la valeur inconnue de p à une valeur de référence p_0.

Test bilatéral d'une proportion

- On veut tester l'hypothèse $H_0 : p = p_0$ contre $H_1 : p \neq p_0$.

Test bilatéral d'une proportion

- On veut tester l'hypothèse $H_0 : p = p_0$ contre $H_1 : p \neq p_0$.
- Si $Z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \in [-z_{1-\frac{\alpha}{2}}; z_{1-\frac{\alpha}{2}}]$, on ne rejette pas H_0.

Test bilatéral d'une proportion

- On veut tester l'hypothèse $H_0 : p = p_0$ contre $H_1 : p \neq p_0$.
- Si $Z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \in [-z_{1-\frac{\alpha}{2}}; z_{1-\frac{\alpha}{2}}]$, on ne rejette pas H_0.
- Sinon on rejette H_0.

Test unilatéral d'une proportion

- On veut tester l'hypothèse $H_0 : p < p_0$ contre $H_1 : p \geq p_0$.

Test unilatéral d'une proportion

- On veut tester l'hypothèse $H_0 : p < p_0$ contre $H_1 : p \geq p_0$.
- Si $Z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq z_{1-\alpha}$, on ne rejette pas H_0.

Test unilatéral d'une proportion

- On veut tester l'hypothèse $H_0 : p < p_0$ contre $H_1 : p \geq p_0$.
- Si $Z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq z_{1-\alpha}$, on ne rejette pas H_0.
- Sinon on rejette H_0.

Statistique bidimensionnelle

Covariance et coefficient de corrélation

- **Covariance empirique :**

$$\sigma_{XY} = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{n} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \mu_X \mu_Y.$$

Covariance et coefficient de corrélation

- **Covariance empirique :**

$$\sigma_{XY} = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{n} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \mu_X \mu_Y.$$

- **Coefficient de corrélation empirique :**

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

Test de linéarité

- $-1 \leq \rho_{XY} \leq 1$.

Test de linéarité

- $-1 \leq \rho_{XY} \leq 1$.
- $|\rho_{XY}| = 1$ si et seulement si tous les points (x_i, y_i) sont parfaitement alignés.

Test de linéarité

- $-1 \leq \rho_{XY} \leq 1$.
- $|\rho_{XY}| = 1$ si et seulement si tous les points (x_i, y_i) sont parfaitement alignés.
- si le coefficient de corrélation est proche de 1 en valeur absolue on pourra espérer une relation linéaire entre X et Y ; on pourra rejeter cette hypothèse dans le cas contraire.

Droite d'ajustement

On cherche donc la droite $Y = aX + b$ "la plus proche possible" de notre nuage de points par la méthode des moindres carrés, c'est-à-dire celle qui minimise la quantité suivante :

$$D(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Théorème

Il existe une et une seule droite qui minimise l'expression $D(a, b)$, cette droite d'équation $y = ax + b$ passe par le point moyen $(\bar{X}, \bar{Y})$ et a pour pente

$$a = \frac{\sigma_{XY}}{\sigma_X^2} = \rho_{XY} \frac{\sigma_Y}{\sigma_X}$$

Il s'agit de la droite de régression de Y par rapport à X.

Exemple

Exemple

On cherche à savoir s'il y a une corrélation linéaire entre le nombre de machines à laver et le nombre de déficients visuels dans une population. Pour cela, on a relevé les échantillons suivants dans un pays d'Europe :

Année	70	71	72	73	74	75
machines (en centaine de milliers)	13	20	23	25	27	31
déficients visuels / 1000 habitants	8	8	9	10	11	11

Année	76	77	78	79	80	81	82	83
machines (en c. de m.)	36	46	55	63	70	76	81	85
déficients visuels $\cdots$	12	16	18	19	20	21	22	23

1. Calculer le coefficient de corrélation de ces échantillons.
2. Que ce coefficient semble-t-il indiquer ?
   Cela vous semble-t-il cohérent ?

Références

Walter Apple
Probabilités pour les non-probabilistes
H&K, édition, 2013

Clément Rau
http ://www.math.univ-toulouse.fr/ rau/
Communication privée, 2015