

HLMA408: Traitement des données

Intervalles de confiance et tests

Joseph Salmon

<http://josephsalmon.eu>

Université de Montpellier



Sommaire

Moyenne empirique d'un échantillon gaussien

Inférence sur la moyenne d'un échantillon gaussien

Inférence sur la moyenne de deux échantillons indépendants

Inférence sur la moyenne de deux échantillons appariés

Inférence sur l'égalité de variance

Modèle gaussien *i.i.d.*

Échantillon: x_1, \dots, x_n (n observations)

Hypothèses sur le modèle:

- ▶ **indépendance** : les x_i suivent des lois indépendantes (pas d'information mutuelle)
- ▶ **identiquement distribuées**: les x_i suivent la même loi
- ▶ la loi de chaque x_i est une loi normale (ou gaussienne) d'espérance μ et de variance σ^2 est inconnu

Rem: on parle alors de modèle *i.i.d.* normal ou gaussien

Rem: au moins l'un des deux paramètres μ ou σ^2 est inconnu

Différents problèmes: un échantillon, une variable

- ▶ Intervalle de confiance sur l'espérance
- ▶ Test d'hypothèse:

$$\mathcal{H}_0 : \mu = \mu_0$$

vs.

$$\mathcal{H}_1 : \mu \neq \mu_0 \text{ (test bilatéral)}$$

$$\mathcal{H}_1 : \mu > \mu_0 \text{ (test unilatéral à droite)}$$

$$\mathcal{H}_1 : \mu < \mu_0 \text{ (test unilatéral à gauche)}$$

Rem: μ_0 est le paramètre de centrage de la loi sous \mathcal{H}_0

Exemple: 1 échantillon, 1 variable

- ▶ Nombre moyen d'une certaine bactérie par unité de volume dans l'eau d'un lac, comparaison avec le niveau "critique" fixé par l'OMS. Si ce niveau est 200

$$\mathcal{H}_0 : \mu = 200 \quad vs. \quad \mathcal{H}_1 : \mu < 200$$

- ▶ Comparaison de la surface d'exploitation agricole récente avec la moyenne (bien connue) de la surface d'une exploitation.
- ▶ Intervalle de l'acidité (pH) de l'eau de pluie.

Différents problèmes: deux échantillons sur deux populations distinctes, une variable

- ▶ Intervalles de confiance sur les espérances ou leurs différences: μ_1 est le paramètre d'intérêt de la population 1, μ_2 celui de la population 2.
- ▶ Test d'hypothèses

$$\mathcal{H}_0 : \mu_1 = \mu_2$$

vs.


$$\mathcal{H}_1 : \mu_1 \neq \mu_2 \text{ (test bilatéral)}$$

$$\mathcal{H}_1 : \mu_1 > \mu_2 \text{ (test unilatéral)}$$

$$\mathcal{H}_1 : \mu_1 < \mu_2 \text{ (test unilatéral)}$$

Exemple: 2 échantillons indépendants, 1 variable

- ▶ Comparaison d'un paramètre physiologique entre un échantillon ayant subi un traitement et un échantillon témoin (**placebo**)
- ▶ Comparaison d'un paramètre physiologique entre mâles et femelles d'une espèce (taille, poids, etc.)
- ▶ etc.

Rem: Au delà de deux groupes, utiliser l'analyse de variance
( : *Analysis of variance, ANOVA*)

1 échantillon, 2 variables mesurées sur chacun des individus

- ▶ Intervalles de confiance sur les espérances ou leurs différences
- ▶ Test d'hypothèses

$$\mathcal{H}_0 : \mu_1 = \mu_2$$

vs.

$$\mathcal{H}_1 : \mu_1 \neq \mu_2 \text{ (test bilatéral)}$$

$$\mathcal{H}_1 : \mu_1 > \mu_2 \text{ (test unilatéral)}$$

$$\mathcal{H}_1 : \mu_1 < \mu_2 \text{ (test unilatéral)}$$

Exemple: 1 échantillon, 2 variables

- ▶ Mesure d'une variable avant et après un traitement, e.g., tester si cette variable a augmenté/diminué,...
- ▶ Mesure de la même variable à des moments différents
- ▶ Mesure de la même variable avec deux appareils de mesure, deux méthodes : température du corps avec un thermomètre électronique ou à mercure
- ▶ Quantité moyenne de masse perdue après un nouveau régime qu'une entreprise cherche à vendre
- ▶ etc.

Sommaire

Moyenne empirique d'un échantillon gaussien

Inférence sur la moyenne d'un échantillon gaussien

Inférence sur la moyenne de deux échantillons indépendants

Inférence sur la moyenne de deux échantillons appariés

Inférence sur l'égalité de variance

Moyenne et variance empiriques

Échantillon : x_1, x_2, \dots, x_n de distribution gaussienne $\mathcal{N}(\mu, \sigma^2)$;
 μ et σ^2 inconnus

Estimer l'espérance: moyenne empirique

$$\bar{x}_n = \frac{x_1 + x_2 + \dots + x_n}{n} \xrightarrow[n \rightarrow \infty]{} \mu \quad (\text{loi des grands nombres})$$

Estimer la variance: variance empirique (dé-biaisée ou non)

$$\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \xrightarrow[n \rightarrow \infty]{} \sigma^2 \quad (\text{loi des grands nombres})$$

$$s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \xrightarrow[n \rightarrow \infty]{} \sigma^2 \quad (\text{loi des grands nombres})$$

Distribution de la moyenne empirique

Rappel: l'échantillon est aléatoire, la moyenne empirique l'est aussi

Théorème

Sous les hypothèses de distribution gaussienne,

$$T_n = \frac{\bar{x}_n - \mu}{\hat{\sigma}_n / \sqrt{n}} \sim t(n - 1)$$

suit une **loi de Student**⁽¹⁾ à $n - 1$ degrés de liberté, notée $t(n - 1)$, avec $\hat{\sigma}_n = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$

⁽¹⁾loi introduite de manière anonyme par William Gosset en 1908, alors employé par l'entreprise Guinness

Distribution de la moyenne empirique

Rappel: l'échantillon est aléatoire, la moyenne empirique l'est aussi

Théorème

Sous les hypothèses de distribution gaussienne,

$$T_n = \frac{\bar{x}_n - \mu}{\hat{\sigma}_n / \sqrt{n}} \sim t(n - 1)$$

suit une **loi de Student**⁽¹⁾ à $n - 1$ degrés de liberté, notée $t(n - 1)$, avec $\hat{\sigma}_n = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$

Densité de la loi de Student de paramètre k :

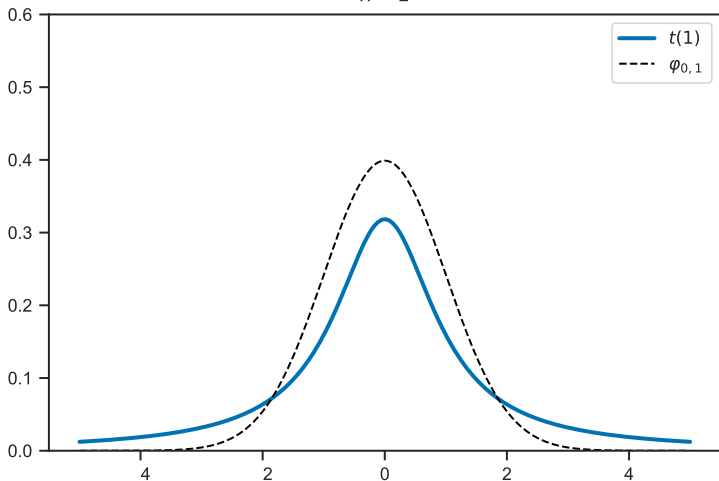
$$f_k(x) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{\pi k} \Gamma(k)} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}$$

⁽¹⁾loi introduite de manière anonyme par William Gosset en 1908, alors employé par l'entreprise Guinness

Loi de Student

Densité d'une loi de student en fonction du nombre de degrés de liberté:

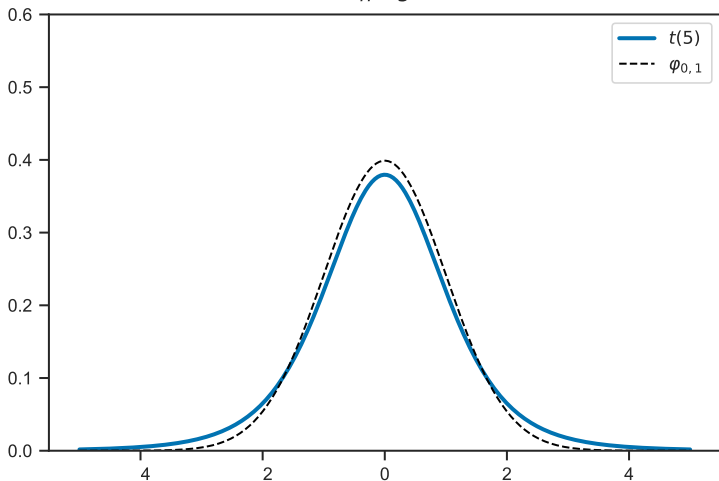
$$n = 1$$



Loi de Student

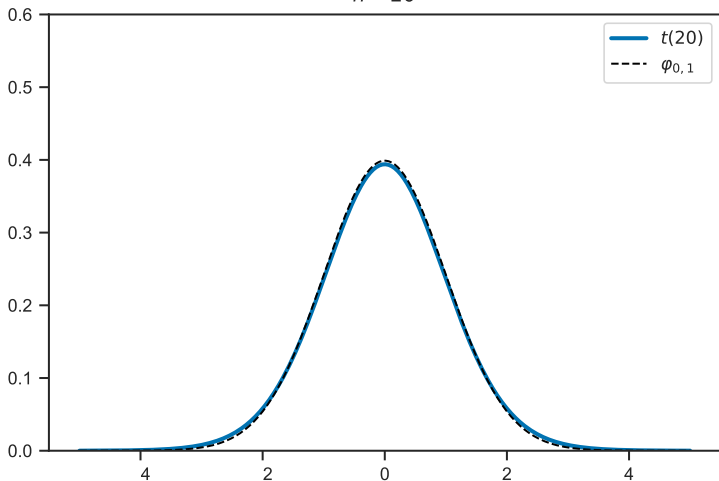
Densité d'une loi de student en fonction du nombre de degrés de liberté:

$$n = 5$$



Loi de Student

Densité d'une loi de student en fonction du nombre de degrés de liberté:
 $n = 20$

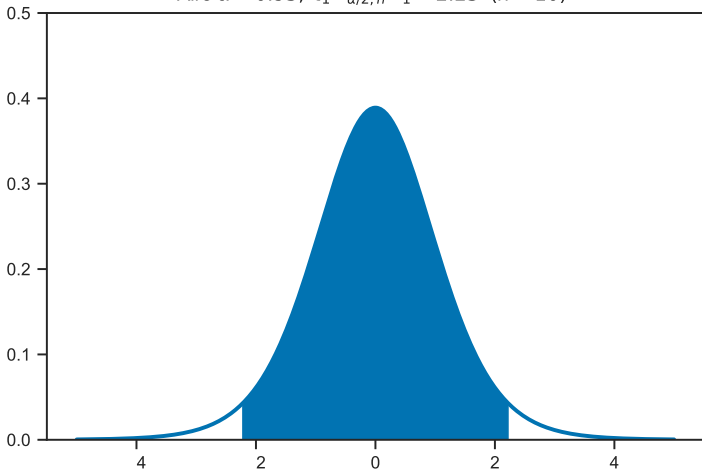


Quantiles et loi de Student

$$\mathbb{P}\left(-t_{1-\frac{\alpha}{2}, n-1} \leq T_n \leq +t_{1-\frac{\alpha}{2}, n-1}\right) = 1 - \alpha \text{ où}$$

$t_{1-\frac{\alpha}{2}, n-1}$: quantile d'ordre $1 - \frac{\alpha}{2}$ d'une loi $t(n-1)$

Aire $\alpha = 0.95$, $t_{1-\alpha/2, n-1} = 2.23$ ($n = 10$)



Sommaire

Moyenne empirique d'un échantillon gaussien

Inférence sur la moyenne d'un échantillon gaussien

Intervalle de confiance

Tests d'hypothèse

Inférence sur la moyenne de deux échantillons indépendants

Inférence sur la moyenne de deux échantillons appariés

Inférence sur l'égalité de variance

Sommaire

Moyenne empirique d'un échantillon gaussien

Inférence sur la moyenne d'un échantillon gaussien

Intervalle de confiance

Tests d'hypothèse

Inférence sur la moyenne de deux échantillons indépendants

Inférence sur la moyenne de deux échantillons appariés

Inférence sur l'égalité de variance

Obtention d'un IC avec la loi de Student

Rappel: $T_n = \frac{\bar{x}_n - \mu}{\hat{\sigma}_n/\sqrt{n}} \sim t(n-1)$

1. $t_{1-\frac{\alpha}{2}, n-1}$: quantile d'ordre $1 - \frac{\alpha}{2}$ d'une loi $t(n-1)$

2. Donc $\mathbb{P}\left(-t_{1-\frac{\alpha}{2}, n-1} \leq T_n \leq +t_{1-\frac{\alpha}{2}, n-1}\right) = 1 - \alpha$.

3. On remplace T_n par son expression

4. Résultat : $\left[\bar{x}_n - t_{1-\frac{\alpha}{2}, n-1} \frac{\hat{\sigma}_n}{\sqrt{n}} ; \bar{x}_n + t_{1-\frac{\alpha}{2}, n-1} \frac{\hat{\sigma}_n}{\sqrt{n}} \right]$
est un IC au niveau $1 - \alpha$ pour le paramètre μ

Rem: calculs identiques au cas IC gaussiens (avec les quantiles de la loi de Student ici, cf. cours précédents)

Propriétés de l'intervalle de confiance (IC)

$$\text{IC: } \left[\bar{x}_n - t_{1-\frac{\alpha}{2}, n-1} \frac{\hat{\sigma}_n}{\sqrt{n}} ; \bar{x}_n + t_{1-\frac{\alpha}{2}, n-1} \frac{\hat{\sigma}_n}{\sqrt{n}} \right]$$

Rem: on contrôle le risque de première espèce inférieur à α

Rem: centré en \bar{x}_n ; longueur $2t_{1-\frac{\alpha}{2}, n-1} \frac{\hat{\sigma}_n}{\sqrt{n}}$

Comme pour le cas gaussien, la longueur de l'IC

- ▶ est proportionnelle à $\hat{\sigma}_n$ ($\approx \sigma$)
- ▶ diminue en \sqrt{n} : il faut multiplier par **100** le nombre d'observations pour gagner un chiffre de précision

Sommaire

Moyenne empirique d'un échantillon gaussien

Inférence sur la moyenne d'un échantillon gaussien

Intervalle de confiance

Tests d'hypothèse

Inférence sur la moyenne de deux échantillons indépendants

Inférence sur la moyenne de deux échantillons appariés

Inférence sur l'égalité de variance

Rappels

Démarche à suivre pour mettre en place un test statistique:

1. Choisir les deux hypothèses : \mathcal{H}_0 et \mathcal{H}_1
2. Choisir/Évaluer une statistique de test : à calculer sur les observations et dont le comportement diffère selon l'hypothèse qui est vraie

Rappels

Démarche à suivre pour mettre en place un test statistique:

1. Choisir les deux hypothèses : \mathcal{H}_0 et \mathcal{H}_1
2. Choisir/Évaluer une statistique de test : à calculer sur les observations et dont le comportement diffère selon l'hypothèse qui est vraie
3. Choisir α : probabilité de rejeter \mathcal{H}_0 à tort que l'on souhaite contrôler (risque de 1^{re} espèce)

Rappels

Démarche à suivre pour mettre en place un test statistique:

1. Choisir les deux hypothèses : \mathcal{H}_0 et \mathcal{H}_1
2. Choisir/Évaluer une statistique de test : à calculer sur les observations et dont le comportement diffère selon l'hypothèse qui est vraie
3. Choisir α : probabilité de rejeter \mathcal{H}_0 à tort que l'on souhaite contrôler (risque de 1^{re} espèce)
4. Déterminer le seuil d'acceptation / rejet associé

Rappels

Démarche à suivre pour mettre en place un test statistique:

1. Choisir les deux hypothèses : \mathcal{H}_0 et \mathcal{H}_1
2. Choisir/Évaluer une statistique de test : à calculer sur les observations et dont le comportement diffère selon l'hypothèse qui est vraie
3. Choisir α : probabilité de rejeter \mathcal{H}_0 à tort que l'on souhaite contrôler (risque de 1^{re} espèce)
4. Déterminer le seuil d'acceptation / rejet associé
5. Conclure rejet / non-rejet de l'hypothèse \mathcal{H}_0

Rappels

Démarche à suivre pour mettre en place un test statistique:

1. Choisir les deux hypothèses : \mathcal{H}_0 et \mathcal{H}_1
2. Choisir/Évaluer une statistique de test : à calculer sur les observations et dont le comportement diffère selon l'hypothèse qui est vraie
3. Choisir α : probabilité de rejeter \mathcal{H}_0 à tort que l'on souhaite contrôler (risque de 1^{re} espèce)
4. Déterminer le seuil d'acceptation / rejet associé
5. Conclure rejet / non-rejet de l'hypothèse \mathcal{H}_0

Choix des hypothèses

En pratique, comment choisir laquelle des deux hypothèses doit être nommée hypothèse nulle \mathcal{H}_0 ?

Plusieurs heuristiques:

- ▶ Choisir comme \mathcal{H}_0 l'hypothèse que l'on cherche à rejeter :
Exemple: test de médicament, \mathcal{H}_0 : “le médicament n'est pas efficace”
Exemple: test de VIH, \mathcal{H}_0 “la personne a le virus”
Exemple: test de grossesse, \mathcal{H}_0 “la femme est enceinte”
- ▶ Si l'une de deux hypothèses est plus simple ou “de dimension plus petite” que l'autre, on la choisit pour \mathcal{H}_0
Exemple: : $\mathcal{H}_0 : \theta = 5$, $\mathcal{H}_1 : \theta \neq 5$
- ▶ Souvent : \mathcal{H}_0 plus “importante” ou plus “dangereuse” que \mathcal{H}_1
Exemple: de la détection de missile \mathcal{H}_0 : “il y a un missile”

Choix des hypothèses (II)

Exemple: tester l'effet d'un médicament avant mise sur le marché
Hypothèse nulle : **toujours** \mathcal{H}_0 "*le médicament est inefficace*"

Conséquences: α contrôle l'erreur suivante

"J'ai annoncé l'efficacité alors qu'il n'a pas d'effet"

Exemple: tester la toxicité d'une quantité de bactéries dans un lac
Hypothèse nulle : \mathcal{H}_0 "*le lac est dangereux*"⁽²⁾

Conséquences: α contrôle l'erreur suivante

"J'ai annoncé que le lac était sain alors qu'il est dangereux"

⁽²⁾imaginons que la limite fixée par l'OMS est 200

Test unilatéral à droite

1. On calcule la valeur observée de $T = \sqrt{n} \frac{\bar{x}_n - \mu_0}{\hat{\sigma}_n}$
Sous \mathcal{H}_0 , T est distribuée suivant $t(n - 1)$
Sous \mathcal{H}_1 , T est négative car $\bar{x}_n - \mu_0 \approx \mu - \mu_0 < 0$
2. Donc rejet de \mathcal{H}_0 lorsque T est plus petit que T_{critique}
3. On choisit T_{critique} pour que $P(t(n - 1) < T_{\text{critique}}) = \alpha$

Concernant la zone de rejet

- sa forme est dictée par le comportement sous \mathcal{H}_1
- ses bornes sont données par le comportement sous \mathcal{H}_0 et par le choix de α .

Test unilatéral à gauche

1. On calcule la valeur observée de $T = \sqrt{n} \frac{\bar{x}_n - \mu_0}{\hat{\sigma}_n}$
Sous \mathcal{H}_0 , T est distribuée suivant $t(n - 1)$
Sous \mathcal{H}_1 , T est **positive** car $\bar{x}_n - \mu_0 \approx \mu - \mu_0 > 0$
2. Donc rejet de \mathcal{H}_0 lorsque T est plus **grand** que T_{critique}
3. On choisit T_{critique} pour que $P(t(n - 1) > T_{\text{critique}}) = \alpha$

Concernant la zone de rejet

- sa forme est dictée par le comportement sous \mathcal{H}_1
- ses bornes sont données par le comportement sous \mathcal{H}_0 et par le choix de α .

Cas bilatéral (III)

1. On calcule la valeur observée de $T = \sqrt{n} \frac{\bar{x}_n - \mu_0}{\hat{\sigma}_n}$
Sous \mathcal{H}_0 , T est distribuée suivant $t(n-1)$
Sous \mathcal{H}_1 , $|T|$ est **grand** car $|\bar{x}_n - \mu_0| \approx |\mu - \mu_0|$ **grand**
2. Donc rejet de \mathcal{H}_0 lorsque $|T|$ est plus **grand** que T_{critique}
3. On choisit T_{critique} pour que $P(|t(n-1)| > T_{\text{critique}}) = \alpha$

Concernant la zone de rejet

- sa forme est dictée par le comportement sous \mathcal{H}_1
- ses bornes sont données par le comportement sous \mathcal{H}_0 et par le choix de α .

Sommaire

Moyenne empirique d'un échantillon gaussien

Inférence sur la moyenne d'un échantillon gaussien

Inférence sur la moyenne de deux échantillons indépendants

Intervalle de confiance pour la différence

Test pour la différence des moyennes

Inférence sur la moyenne de deux échantillons appariés

Inférence sur l'égalité de variance

Sommaire

Moyenne empirique d'un échantillon gaussien

Inférence sur la moyenne d'un échantillon gaussien

Inférence sur la moyenne de deux échantillons indépendants

Intervalle de confiance pour la différence

Test pour la différence des moyennes

Inférence sur la moyenne de deux échantillons appariés

Inférence sur l'égalité de variance

Deux échantillons indépendants

x_1, x_2, \dots, x_{n_1} : échantillon sur la première population de taille n_1

y_1, y_2, \dots, y_{n_2} : échantillon sur la deuxième population de taille n_2

Échantillon 1 : distribution gaussienne $\mathcal{N}(\mu_1, \sigma_1^2)$

Échantillon 2 : distribution gaussienne $\mathcal{N}(\mu_2, \sigma_2^2)$

Objectif: obtenir des résultats sur $\mu_1 - \mu_2$

Notation:

$$\hat{\sigma}_{n_1}(x)^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x}_{n_1})^2$$

$$\hat{\sigma}_{n_2}(y)^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_i - \bar{y}_{n_2})^2$$

Exemple: reprendre les données sur la taille des parents

Variances égales: cas $\sigma_1 = \sigma_2$

Intervalle de confiance, variances égales⁽³⁾

Un IC au niveau $(1 - \alpha)$ pour $\mu_1 - \mu_2$ est

$$(\bar{x}_{n_1} - \bar{y}_{n_2}) \pm t_{1-\frac{\alpha}{2}, n_1+n_2-2} \hat{\sigma}_{\text{groupé}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \quad \text{où}$$

$$\hat{\sigma}_{\text{groupé}}^2 = \frac{(n_1 - 1)\hat{\sigma}_{n_1}(x)^2 + (n_2 - 1)\hat{\sigma}_{n_2}(y)^2}{n_1 + n_2 - 2} \text{ et}$$

$t_{1-\frac{\alpha}{2}, n_1+n_2-2}$: quantile $(1 - \frac{\alpha}{2})$ d'une Student $t(n_1 + n_2 - 2)$

Rem:

$$\begin{aligned} \sum_{i=1}^{n_1} (x_i - \bar{x}_{n_1})^2 + \sum_{i=1}^{n_2} (y_i - \bar{y}_{n_2})^2 &\sim \chi^2(n_1 - 1) + \chi^2(n_2 - 1) \\ &\sim \chi^2(n_1 + n_2 - 2) \quad (\text{indépendance}) \end{aligned}$$

⁽³⁾ preuve p. 148, poly http://josephsalmon.eu/enseignement/ENSAE/StatAppli_tsybakov.pdf

Variances différentes: σ_1 et σ_2 quelconques

Intervalle de confiance, variances distinctes⁽⁴⁾

Un IC au niveau $(1 - \alpha)$ pour $\mu_1 - \mu_2$ est

$$(\bar{x}_{n_1} - \bar{y}_{n_2}) \pm t_{1-\frac{\alpha}{2}, k} \sqrt{\frac{\hat{\sigma}_{n_1}(x)^2}{n_1} + \frac{\hat{\sigma}_{n_2}(y)^2}{n_2}}, \quad \text{où}$$

où $k = \min(n_1 - 1, n_2 - 1)$

Rem: cet intervalle de confiance est approximatif, plus simplement on peut aussi considérer un IC basé sur les quantiles gaussiens...

⁽⁴⁾P. J. Bickel and K. A. Doksum. *Mathematical statistics*. Basic ideas and selected topics, Holden-Day Series in Probability and Statistics. San Francisco, Calif.: Holden-Day Inc., 1976.

Variances différentes: σ_1 et σ_2 quelconques (asymptotique)

Intervalle de confiance, variances différentes

Un IC asymptotique au niveau $(1 - \alpha)$ pour $\mu_1 - \mu_2$ est

$$(\bar{x}_{n_1} - \bar{y}_{n_2}) \pm q_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}_{n_1}^2(x)}{n_1} + \frac{\hat{\sigma}_{n_2}^2(y)}{n_2}}, \quad \text{où}$$

$q_{1-\frac{\alpha}{2}}$: quantile $(1 - \frac{\alpha}{2})$ d'une loi normale centrée réduite

Sommaire

Moyenne empirique d'un échantillon gaussien

Inférence sur la moyenne d'un échantillon gaussien

Inférence sur la moyenne de deux échantillons indépendants

Intervalle de confiance pour la différence

Test pour la différence des moyennes

Inférence sur la moyenne de deux échantillons appariés

Inférence sur l'égalité de variance


Test, variance égale

On veut tester $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ où δ_0 est fixé par l'utilisateur

Test, si on suppose $\sigma_1 = \sigma_2$

$$T = \frac{(\bar{x}_{n_1} - \bar{y}_{n_2}) - \delta_0}{\hat{\sigma}_{\text{groupé}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{avec} \quad df = n_1 + n_2 - 2$$

Alternative	Zone de rejet	$t(df)$ -quantile d'ordre
$\mathcal{H}_1 : \mu_1 - \mu_2 > \delta_0$	$T \geq t_*$	$1 - \alpha$
$\mathcal{H}_1 : \mu_1 - \mu_2 < \delta_0$	$T \leq t_{**}$	α
$\mathcal{H}_1 : \mu_1 - \mu_2 \neq \delta_0$	$ T \geq t_{***}$	$1 - \frac{\alpha}{2}$

Rem: df est l'acronyme de degré de liberté ( : *degree of freedom*)


Test, sans hypothèse de variance

On veut tester $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ où δ_0 est fixé par l'utilisateur.

Test, sans hypothèse sur σ_1 et σ_2

$$T = \frac{(\bar{x}_{n_1} - \bar{y}_{n_2}) - \delta_0}{\sqrt{\frac{\hat{\sigma}_{n_1}^2}{n_1} + \frac{\hat{\sigma}_{n_2}^2}{n_2}}} \quad \text{avec} \quad df = \min(n_1 - 1, n_2 - 1)$$

Alternative	Zone de rejet	$t(df)$ -quantile d'ordre
$\mathcal{H}_1 : \mu_1 - \mu_2 > \delta_0$	$T \geq t_*$	$1 - \alpha$
$\mathcal{H}_1 : \mu_1 - \mu_2 < \delta_0$	$T \leq t_{**}$	α
$\mathcal{H}_1 : \mu_1 - \mu_2 \neq \delta_0$	$ T \geq t_{***}$	$1 - \frac{\alpha}{2}$

Rem: df est l'acronyme de degré de liberté ( : *degree of freedom*)

Sommaire

Moyenne empirique d'un échantillon gaussien

Inférence sur la moyenne d'un échantillon gaussien

Inférence sur la moyenne de deux échantillons indépendants

Inférence sur la moyenne de deux échantillons appariés

Intervalle de confiance pour la différence

Test pour la différence des moyennes

Inférence sur l'égalité de variance

Sommaire

Moyenne empirique d'un échantillon gaussien

Inférence sur la moyenne d'un échantillon gaussien

Inférence sur la moyenne de deux échantillons indépendants

Inférence sur la moyenne de deux échantillons appariés

Intervalle de confiance pour la différence

Test pour la différence des moyennes

Inférence sur l'égalité de variance

Cadre

Objectif: comparer les moyennes de deux séries de mesures faites sur les mêmes unités statistiques

Cas 1 : x_1, x_2, \dots, x_n : premier échantillon, de distribution gaussienne $\mathcal{N}(\mu_1, \sigma_1^2)$

Cas 2 : y_1, y_2, \dots, y_n : second échantillon, de distribution gaussienne $\mathcal{N}(\mu_2, \sigma_2^2)$

Pour le i^e individu, on pose $d_i = x_i - y_i$

Exemple: la capacité photosynthétique est suivie sur 30 plantes à deux moments de la journée (matin et après-midi)

Résumés sur la variable D

Moyenne empirique de la différence

$$\bar{d}_n = \frac{1}{n} \sum_{i=1}^n d_i \quad (\approx \mu_1 - \mu_2)$$

Variance empirique de la différence

$$\hat{\sigma}_d^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d}_n)^2$$

Intervalle de confiance, variances quelconques

Intervalle de confiance

Un IC au niveau $(1 - \alpha)$ pour $\mu_1 - \mu_2$ est

$$\bar{d}_n \pm t_{1-\frac{\alpha}{2}, n-1} \frac{\hat{\sigma}_d}{\sqrt{n}}$$

Sommaire

Moyenne empirique d'un échantillon gaussien

Inférence sur la moyenne d'un échantillon gaussien

Inférence sur la moyenne de deux échantillons indépendants

Inférence sur la moyenne de deux échantillons appariés

Intervalle de confiance pour la différence

Test pour la différence des moyennes

Inférence sur l'égalité de variance

Test

On veut tester $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ où δ_0 est fixé par l'utilisateur

Test

$$T = \frac{\bar{d}_n - \delta_0}{\hat{\sigma}_d / \sqrt{n}} \text{ avec } df = n - 1$$

Hypothèse alternative	Zone de rejet	$t(df)$ -quantile d'ordre
$\mathcal{H}_1 : \mu_1 - \mu_2 > \delta_0$	$T \geq t_*$	$1 - \alpha$
$\mathcal{H}_1 : \mu_1 - \mu_2 < \delta_0$	$T \leq t_{**}$	α
$\mathcal{H}_1 : \mu_1 - \mu_2 \neq \delta_0$	$ T \geq t_{***}$	$1 - \frac{\alpha}{2}$

Sommaire

Moyenne empirique d'un échantillon gaussien

Inférence sur la moyenne d'un échantillon gaussien

Inférence sur la moyenne de deux échantillons indépendants

Inférence sur la moyenne de deux échantillons appariés

Inférence sur l'égalité de variance

Deux échantillons indépendants

x_1, x_2, \dots, x_{n_1} : échantillon sur la première population de taille n_1

y_1, y_2, \dots, y_{n_2} : échantillon sur la deuxième population de taille n_2

Échantillon 1 : distribution gaussienne $\mathcal{N}(\mu_1, \sigma_1^2)$

Échantillon 2 : distribution gaussienne $\mathcal{N}(\mu_2, \sigma_2^2)$

Objectif: obtenir des résultats sur $\mu_1 - \mu_2$

Notation:

$$\hat{\sigma}_{n_1}(x)^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x}_{n_1})^2$$

$$\hat{\sigma}_{n_2}(y)^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_i - \bar{y}_{n_2})^2$$

Test d'égalité de variance

On cherche à tester: \mathcal{H}_0 : " $\sigma_1^2 = \sigma_2^2$ " vs. \mathcal{H}_1 : " $\sigma_1^2 \neq \sigma_2^2$ "

Une statistique classique pour cela est :

$$\frac{\hat{\sigma}_{n_1}(x)^2}{\hat{\sigma}_{n_2}(y)^2} = \frac{\frac{1}{n_1-1} \sum_{i=1}^{n_1} (x_i - \bar{x}_{n_1})^2}{\frac{1}{n_2-1} \sum_{i=1}^{n_2} (y_i - \bar{y}_{n_2})^2}$$

Sous l'hypothèse \mathcal{H}_0 cette statistique suit une loi de Fisher (ou de Fisher-Snedecor) de paramètres $(n_1 - 1, n_2 - 1)$ ce que l'on note

$$\frac{\hat{\sigma}_{n_1}(x)^2}{\hat{\sigma}_{n_2}(y)^2} \sim \mathcal{F}(n_1 - 1, n_2 - 1)$$

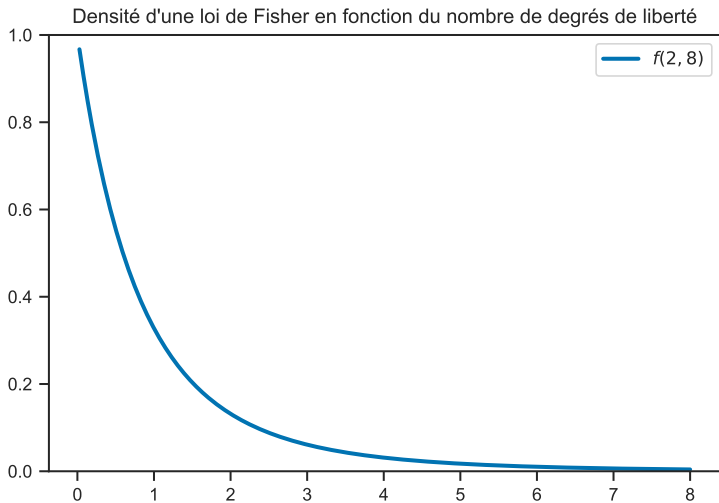
Loi de Fisher

La densité d'une loi de Fisher, notée $\mathcal{F}(d_1, d_2)$, est donnée par

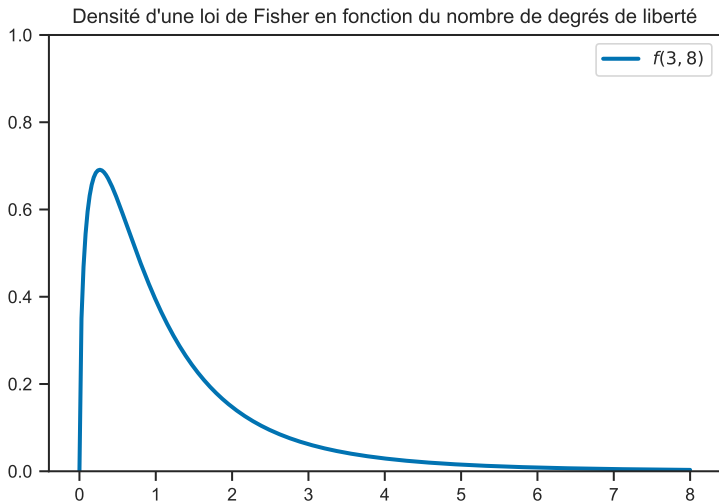
$$f(x) = \frac{\left(\frac{xd_1}{xd_1+d_2}\right)^{\frac{d_1}{2}} \left(1 - \frac{xd_1}{xd_1+d_2}\right)^{\frac{d_2}{2}}}{xB\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

pour tout réel $x \geq 0$, où d_1 et d_2 sont des entiers positifs et B est la fonction bêta

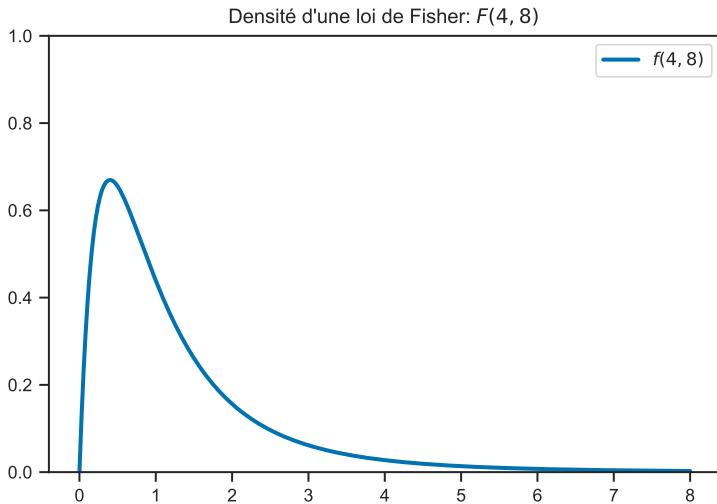
Densité



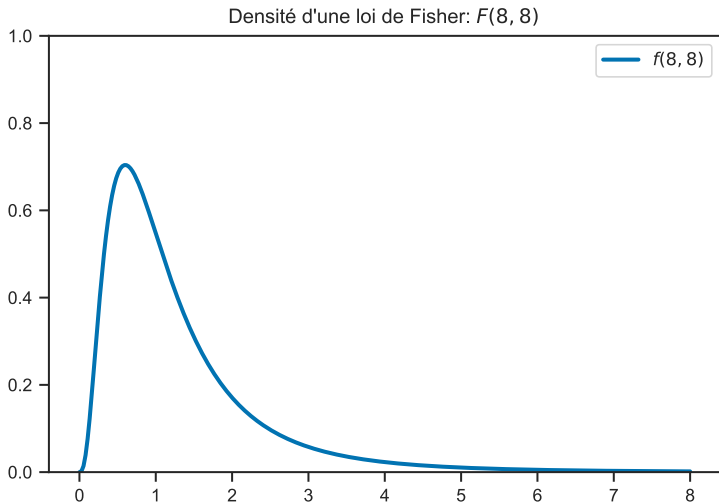
Densité



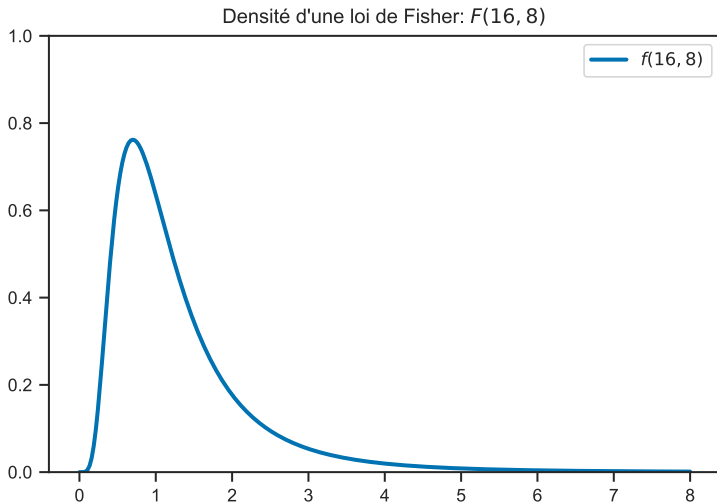
Densité



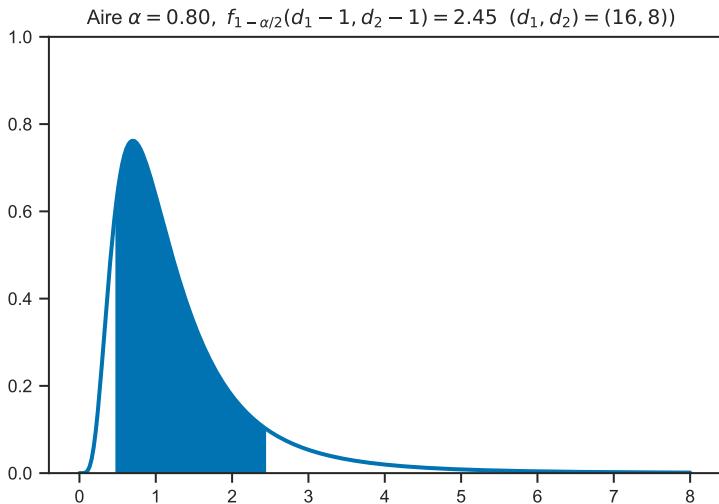
Densité



Densité



Quantile



Propriété

La loi de Fisher peut être construite comme le quotient de deux variables aléatoires indépendantes, U_1 et U_2 , distribuées des lois du χ^2 de degrés de liberté d_1 (resp. d_2) :

$$\mathcal{F}(d_1, d_2) \sim \frac{U_1/d_1}{U_2/d_2}$$

Bibliographie I

- ▶ Bickel, P. J. and K. A. Doksum. *Mathematical statistics*. Basic ideas and selected topics, Holden-Day Series in Probability and Statistics. San Francisco, Calif.: Holden-Day Inc., 1976.