HLMA408: Traitement des données

Intervalles de confiance et tests

Joseph Salmon

http://josephsalmon.eu

Université de Montpellier

Sommaire

Modèle gaussien *i.i.d.*

Échantillon: $x_1, \ldots, x_n$ (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

Cas: 1 échantillon, 1 variable

- Intervalle de confiance sur l'espérance
- Tests d'hypothèse :
  $\mathcal{H}_0$: "$\mu = \mu_0$"
  vs.
  $\mathcal{H}_1$: "$\mu \neq \mu_0$" **test bilatéral**
  $\mathcal{H}_1$: "$\mu > \mu_0$", **test unilatéral à droite** (rejet à droite)
  $\mathcal{H}_1$: "$\mu < \mu_0$", **test unilatéral à gauche** (rejet 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 du rendement d'une d'exploitation agricole récente avec le rendement d'une exploitation type.

- Mesure de l'acidité (pH) de l'eau de pluie.

Cas: 2 échantillons indép., 1 variable

- Intervalles de confiance sur les espérances ou leurs différences :
  - μ_1 est le paramètre d'intérêt de la population 1
  - μ_2 est le paramètre d'intérêt de la population 2

- Test d'hypothèses
  $\mathcal{H}_0$: "$\mu_1 = \mu_2$"
  vs.
  $\mathcal{H}_1$: "$\mu_1 \neq \mu_2$", **test bilatéral**
  $\mathcal{H}_1$: "$\mu_1 > \mu_2$", **test unilatéral**
  $\mathcal{H}_1$: "$\mu_1 < \mu_2$", **test unilatéral**

Exemple: 2 échantillons indép., 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, masse, etc.)
- etc.

Rem. : au delà de deux groupes $\Longrightarrow$ analyse de variance (: *Analysis of variance, ANOVA*)

Cas: 1 échantillon, 2 variables mesurées

- Intervalles de confiance sur les espérances ou leurs différences

- Test d'hypothèses
  $\mathcal{H}_0$: "$\mu_1 = \mu_2$"
  vs.
  $\mathcal{H}_1$: "$\mu_1 \neq \mu_2$", **test bilatéral**
  $\mathcal{H}_1$: "$\mu_1 > \mu_2$", **test unilatéral**
  $\mathcal{H}_1$: "$\mu_1 < \mu_2$", **test unilatéral**

Exemple: 1 échantillon, 2 variables

- Mesure d'une variable avant et après un traitement, *e.g.*, tester si cette variable a augmenté ou diminué, etc.
- 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 et variance empiriques

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

Estimer l'espérance : moyenne empirique

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

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

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

et

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

Distribution de la moyenne empirique

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

Théorème

Sous les hypothèses de distribution gaussienne, la statistique

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

suit une **loi de Student**[1] à $n-1$ degrés de liberté, notée $t(n-1)$, où l'on note $\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$.

[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, la statistique

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

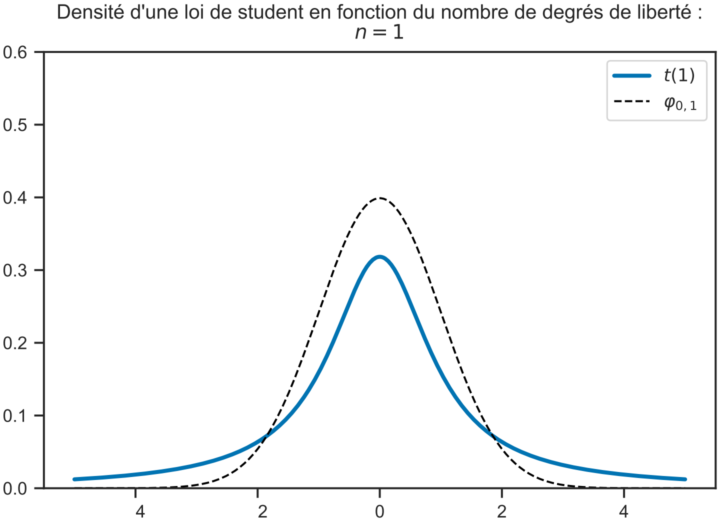
suit une **loi de Student**[1] à $n-1$ degrés de liberté, notée $t(n-1)$, où l'on note $\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$.

f_k : densité de la loi de Student de paramètre k est donnée par

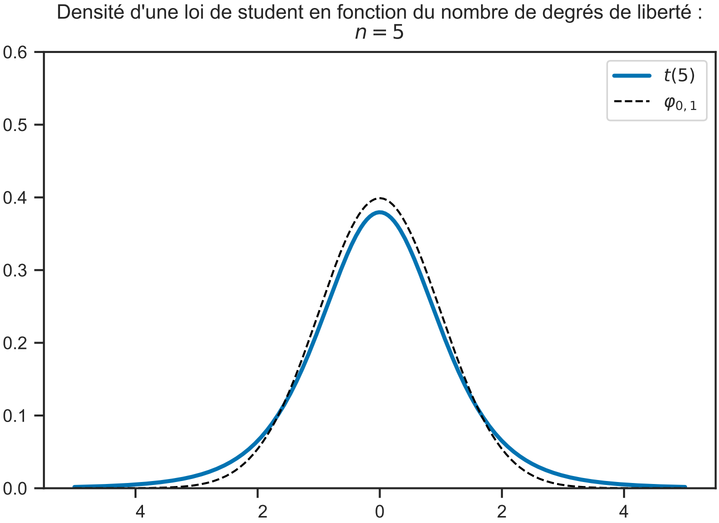
$$f_k(x) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{\pi k}\Gamma(k)}\left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}} \propto \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}$$

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

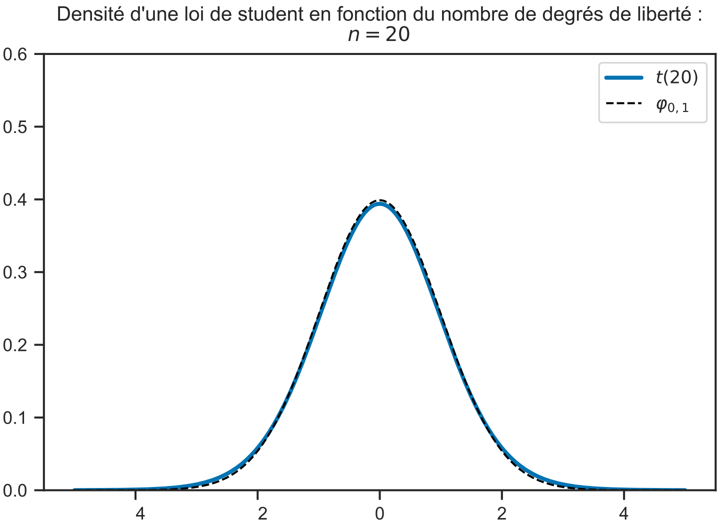
Loi de Student



Loi de Student



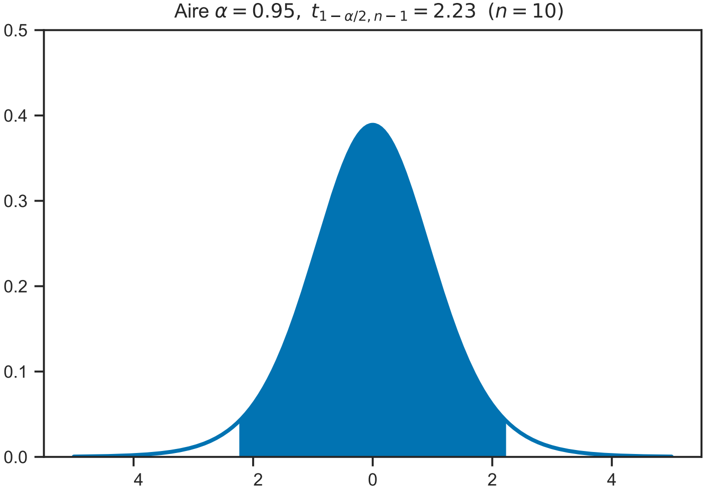
Loi de Student



Quantiles et loi de Student

$$\mathbb{P}\Big(- t_{1-\alpha/2,n-1} \leq T_n \leq +t_{1-\alpha/2,n-1} \Big) = 1 - \alpha$$

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



Sommaire

Sommaire

Obtention d'un IC avec la loi de Student

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

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

Rem. : IC similaire au cas gaussien, mais les quantiles gaussiens sont remplacés par les quantiles de Student, *cf.* cours précédents

Propriétés de l'intervalle de confiance (IC)

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

Rem. : risque de première espèce contrôlé (inférieur à) au niveau α

Rem. : IC 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 diviser par **10** la largeur de l'IC

Sommaire

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 1re espèce)
4. Déterminer le seuil d'acceptation / rejet associé
5. Conclure rejet / non-rejet de l'hypothèse $\mathcal{H}_0$

Test unilatéral à gauche: $\mathcal{H}_1$: "$\mu < \mu_0$"

1. On calcule la valeur observée de $T = \sqrt{n}\left(\dfrac{\bar{x}_n - \mu_0}{\hat{\sigma}_n}\right)$

   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 la loi sous $\mathcal{H}_0$ et par le choix de α

Test unilatéral à droite: $\mathcal{H}_1$: "$\mu > \mu_0$"

1. On calcule la valeur observée de $T = \sqrt{n}\left(\frac{\bar{x}_n - \mu_0}{\hat{\sigma}_n}\right)$
   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 la loi sous $\mathcal{H}_0$ et par le choix de α

Cas bilatéral (III)

1. On calcule la valeur observée de $T = \sqrt{n}\left(\frac{\bar{x}_n - \mu_0}{\hat{\sigma}_n}\right)$
   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 $P\big(|t(n-1)| > T_{\text{critique}}\big) = \alpha$

Concernant la zone de rejet

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

Sommaire

Sommaire

Deux échantillons indépendants

$x_1, x_2, \ldots, x_{n_1}$: échantillon sur la première population de taille n_1
$y_1, y_2, \ldots, y_{n_2}$: 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:

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

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

Exemple : reprendre les données sur la taille des parents

Variances égales: cas $\sigma_1 = \sigma_2$

Intervalle de confiance, variances égales[2] p. 148

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

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

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

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

Rem. :

$$\sum_{i=1}^{n_1}(x_i - \bar{x}_{n_1})^2 + \sum_{i=1}^{n_2}(y_i - \bar{y}_{n_2})^2 \sim \chi^2(n_1-1) + \chi^2(n_2-1)$$

$$\sim \chi^2(n_1+n_2-2) \quad \text{(indépendance)}$$

[2]preuve hors programme, http://josephsalmon.eu/enseignement/ENSAE/StatAppli_tsybakov.pdf

Variances différentes: σ_1 et σ_2 quelconques

Intervalle de confiance, variances distinctes[3]

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

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

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...

[3] 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 (IC asymptotique)

Intervalle de confiance, variances différentes

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

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

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

Sommaire

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 $\sigma_1 = \sigma_2$

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

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. : **degré de liberté** (🇬🇧 : *degree of freedom*); abréviation *df*

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 supposer que $\sigma_1 = \sigma_2$

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

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
$\mathcal{H}_1 \ : \ \mu_1 - \mu_2 \neq \delta_0$	$\lvert T\rvert \geq t_{***}$	$1 - \frac{\alpha}{2}$

Rem. : **degré de liberté** (: *degree of freedom*); abréviation *df*

Sommaire

Sommaire

Cadre

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

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

Cas 2 : $y_1, y_2, \ldots, y_n$: 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 : capacité photosynthétique suivie sur 30 plantes, à deux moments de la journée (matin et après-midi)

Résumés sur la variable D

Moyenne empirique de la différence :

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

Variance empirique de la différence :

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

Intervalle de confiance, variances quelconques

Intervalle de confiance

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

$$\left[\bar{d}_n - t_{1-\frac{\alpha}{2},n-1}\frac{\hat{\sigma}_{\mathsf{diff}}}{\sqrt{n}}\ ;\ \bar{d}_n + t_{1-\frac{\alpha}{2},n-1}\frac{\hat{\sigma}_{\mathsf{diff}}}{\sqrt{n}}\right] \quad .$$

Sommaire

Test

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

Test

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

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
$\mathcal{H}_1 \; : \; \mu_1 - \mu_2 \neq \delta_0$	$\lvert T \rvert \geq t_{***}$	$1 - \frac{\alpha}{2}$

Sommaire

Deux échantillons indépendants

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

$y_1, y_2, \ldots, y_{n_2}$: 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 :

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

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

Test d'égalité de variance

On cherche à tester :

$$\mathcal{H}_0 : \text{``}\sigma_1^2 = \sigma_2^2\text{''} \quad \text{vs.} \quad \mathcal{H}_1 : \text{``}\sigma_1^2 \neq \sigma_2^2\text{''}$$

Une statistique classique pour cela est :

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

Sous l'hypothèse $\mathcal{H}_0$ cette statistique suit une loi de Fisher[4] (ou de Fisher-Snedecor) de paramètres $(n_1 - 1, n_2 - 1)$, ce que l'on note

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

[4] https://fr.wikipedia.org/wiki/Loi_de_Fisher

Loi de Fisher (formule hors-programme)

Densité d'une loi de Fisher,[5] notée $\mathcal{F}(d_1, d_2)$:

$$\forall x \geq 0, \quad f(x) = \frac{\left(\frac{xd_1}{xd_1+d_2}\right)^{\frac{d_1}{2}} \left(1 - \frac{xd_1}{xd_1+d_2}\right)^{\frac{d_2}{2}}}{x\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

d_1, d_2 : entiers positifs ; $\quad B$: la fonction bêta[6]

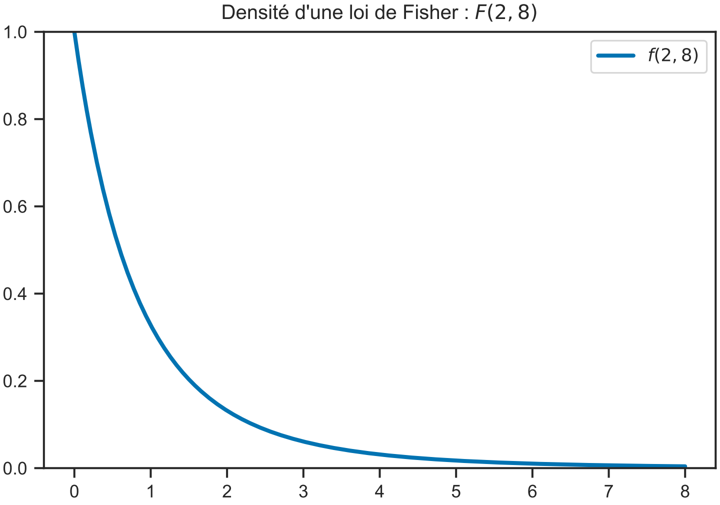
Rem. : la loi de Fisher se construit comme le quotient de deux variables aléatoires indépendantes, U_1 et U_2, distribuées selon des lois du χ^2 de degrés de liberté d_1 (resp. d_2) :

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

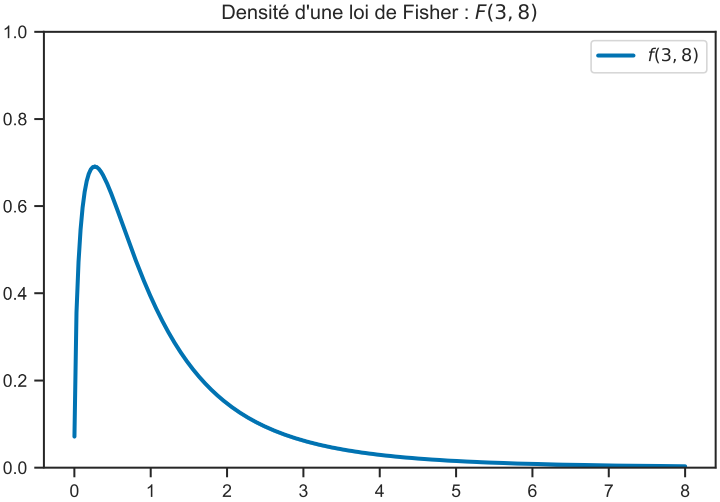
[5] https://fr.wikipedia.org/wiki/Loi_de_Fisher

[6] https://fr.wikipedia.org/wiki/Fonction_bêta

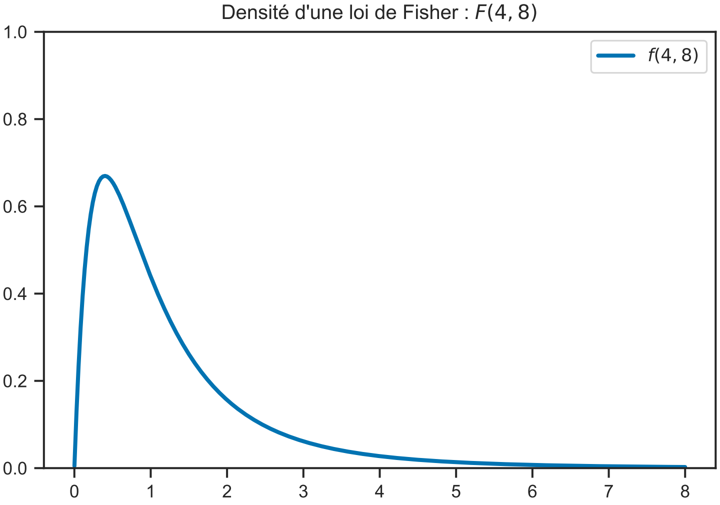
Densité



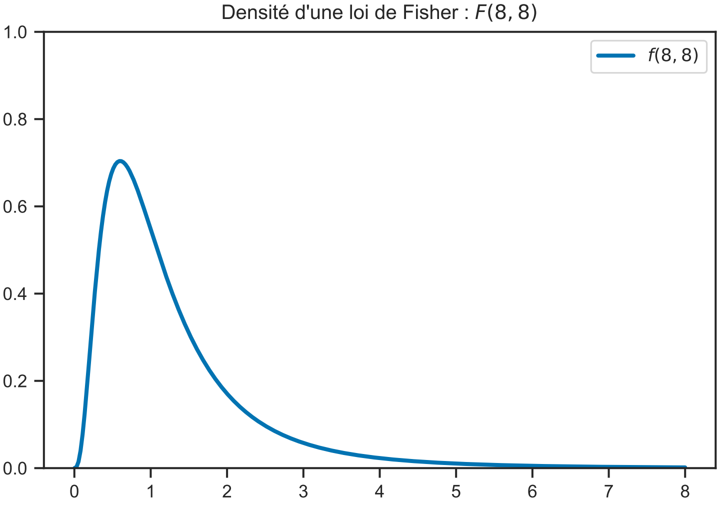
Densité



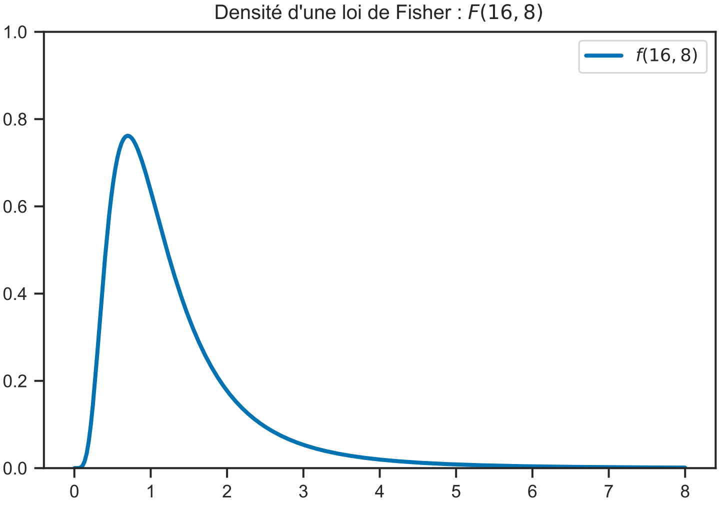
Densité



Densité

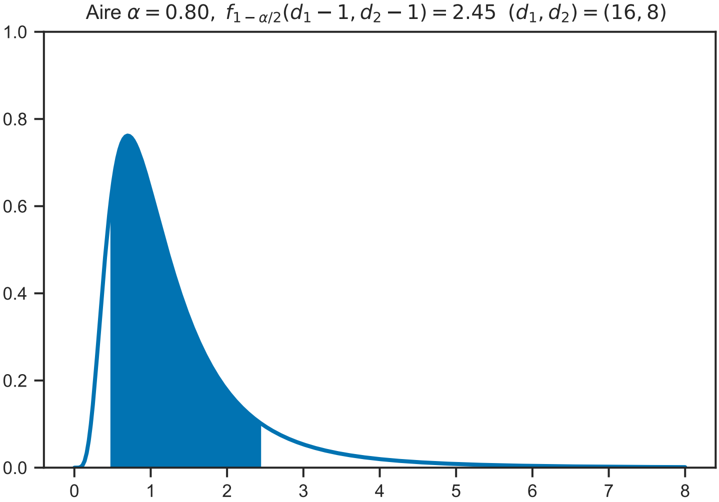


Densité



Densité d'une loi de Fisher : $F(16, 8)$
$f(16, 8)$
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8

Quantile



Aire $\alpha = 0.80,\ f_{1-\alpha/2}(d_1-1, d_2-1) = 2.45\ \ (d_1, d_2) = (16, 8)$
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8

Quantiles d'une loi de Fisher (en Python)

```
>>> from scipy.stats import f
>>> # degre de liberté numérateur/dénominateur
>>> dfn, dfd = 4, 5
>>> quantile = 1 - 0.05
>>> f.ppf(quantile, dfn, dfd)
5.192167772803923
```

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.