Inférence statistique

Séance 7

Tests statistiques d'hypothèse:

rappels
comparaison de deux échantillons gaussiens
tests d'ajustement à une loi
tests d'indépendance

Frédéric Sur

Mines Nancy

https://members.loria.fr/FSur/

Plan

Loi d'une variable aléatoire

Variable aléatoire discrète (à valeurs dans $\mathbb{N}$)
loi de X = suite $p_n = P(X = n)$

Variable aléatoire continue (à valeurs dans $\mathbb{R}$)
loi de X = les $P(X \in I)$ pour tout I borélien de $\mathbb{R}$.

$\rightarrow$ Variable aléatoire absolument continue
loi de la v.a. déterminée par sa *densité* : $f > 0$ intégrable t.q. $\int_R f = 1$, et pour tout intervalle $I \subset \mathbb{R}$,

$$P(X \in I) = \int_I f(x)\,\mathrm{d}x$$

Lorsque les intégrales existent :
$E(X) = \int_{\mathbb{R}} xf(x)\,\mathrm{d}x$ et $\mathrm{Var}(X) = \int_{\mathbb{R}} (x - E(X))^2 f(x)\,\mathrm{d}x$

Fonction de répartition

Soit X une variable aléatoire.
Sa **fonction de répartition** F est définie par :

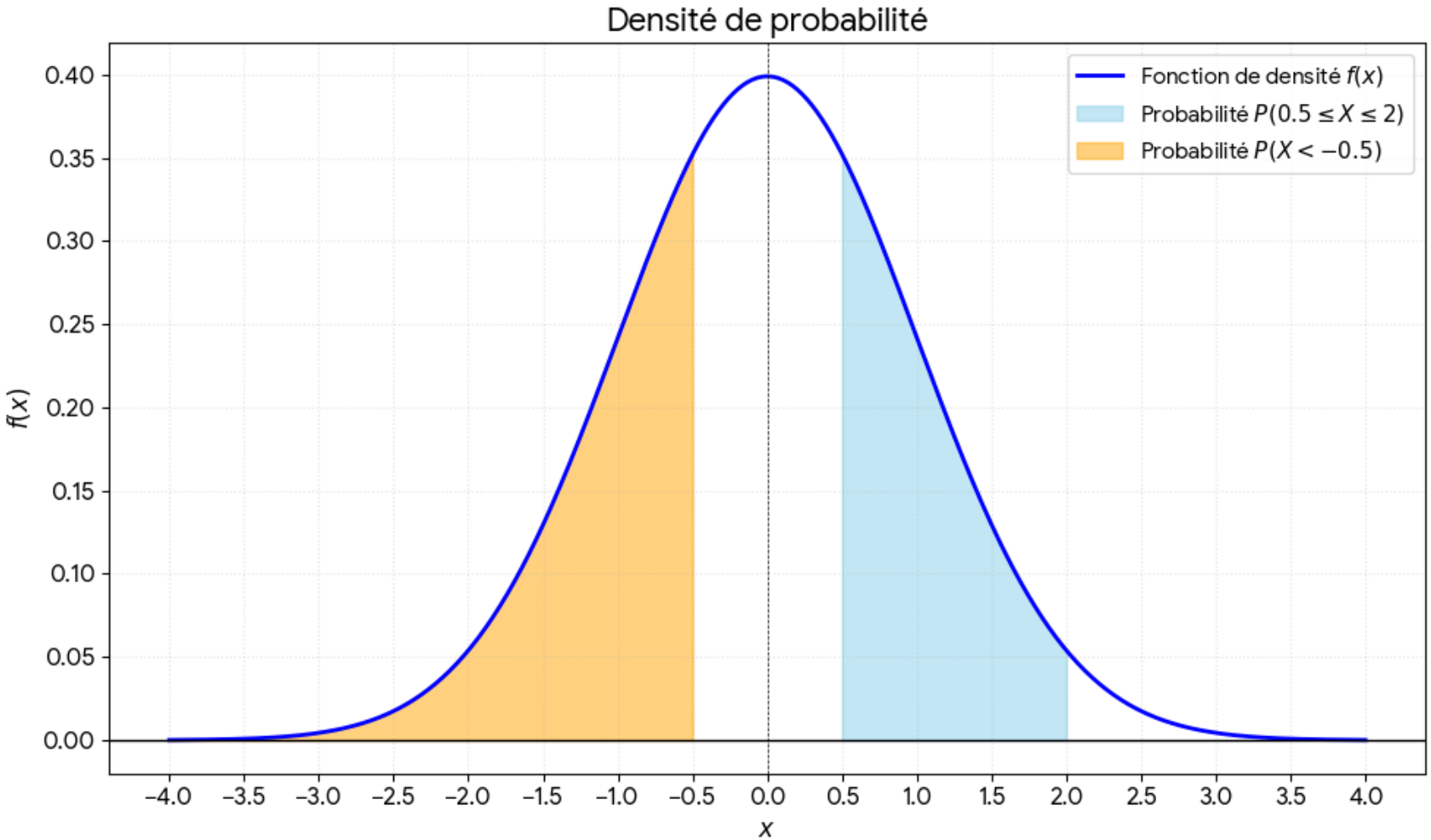
$$F : \begin{cases} \mathbb{R} \to \mathbb{R} \\ x \mapsto P(X \leqslant x) \quad = \int_{-\infty}^{x} f(x)\,\mathrm{d}x \text{ si v.a. a.c.} \end{cases}$$

Propriétés :

- pour tout $x \in \mathbb{R}$, $0 \leqslant F(x) \leqslant 1$
- F est croissante
- $\lim_{x \to -\infty} F(x) = 0$ et $\lim_{x \to +\infty} F(x) = 1$
- en tout $x \in \mathbb{R}$ où f est continue, $F'(x) = f(x)$.

Fonction de répartition et densité

Pour une v.a. absolument continue :



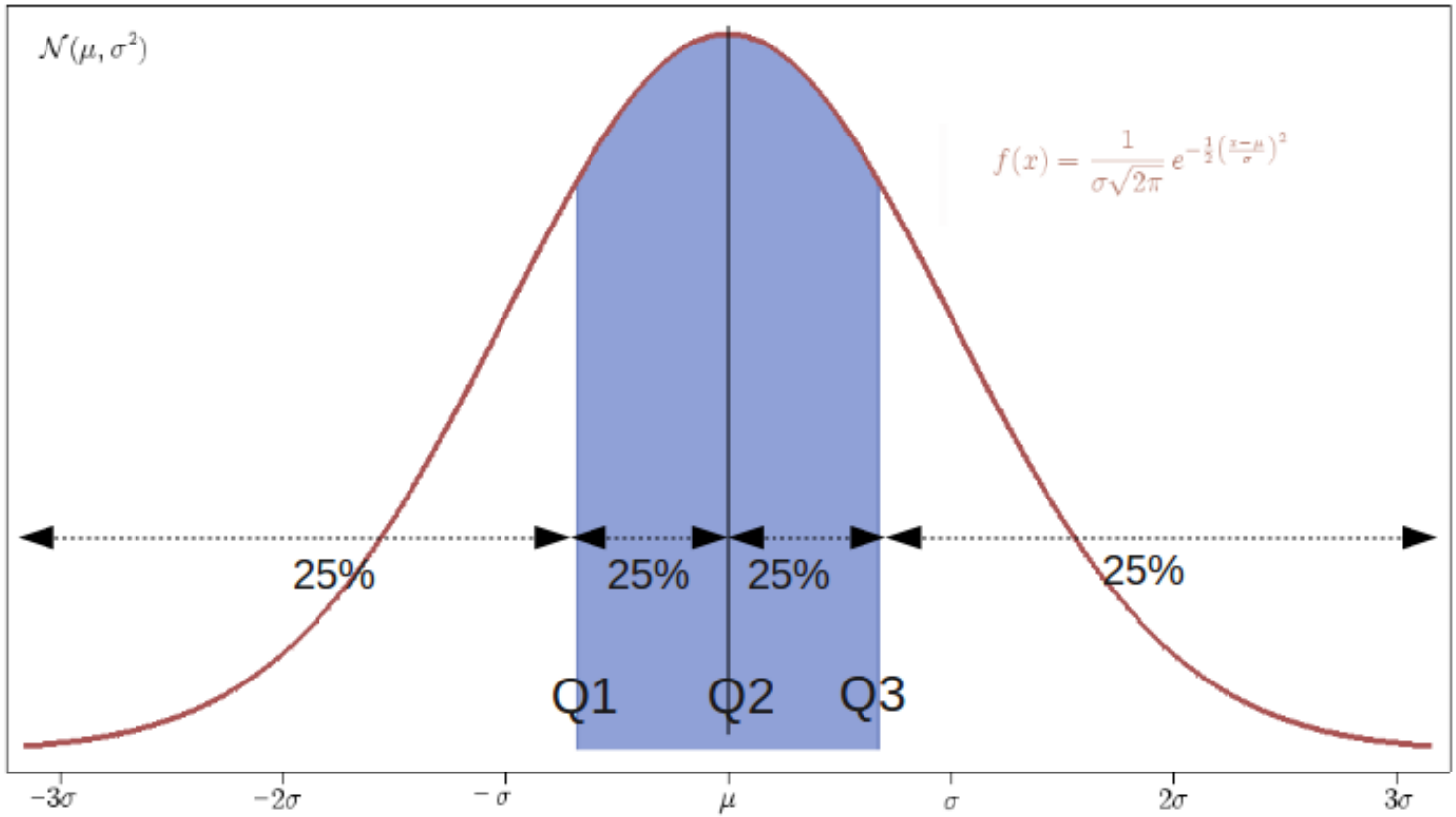
$$P(X < -0,5) = F(-0,5), \quad P(X > -0,5) = 1 - F(-0,5)$$

$$P(0,5 \leqslant X \leqslant 2) = F(2) - F(0,5)$$

Quantiles

Quantiles : valeurs qui divisent un jeu de données en intervalles de même probabilité

Exemples : *quartiles*, médiane



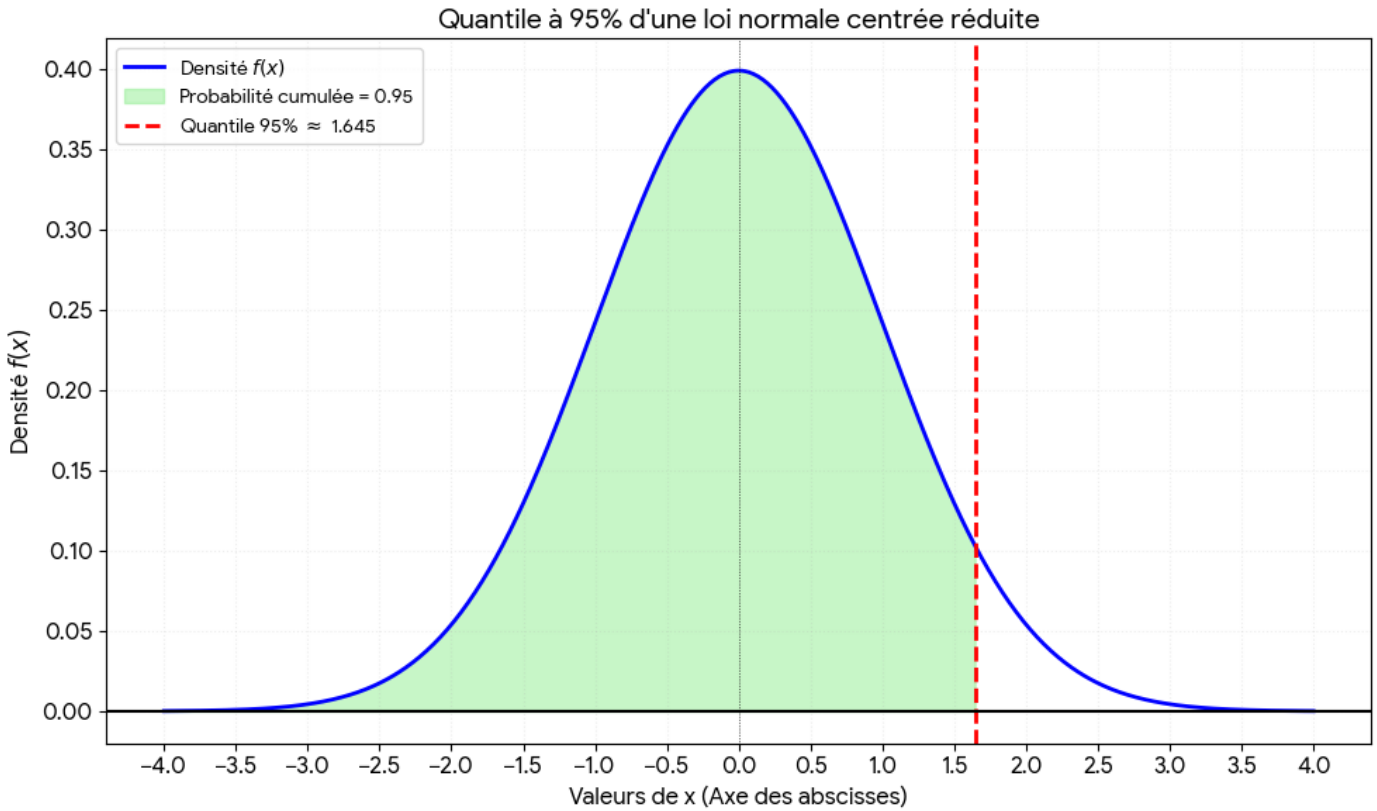
Par Iqr.png : Ark0nderivative work : Gato ocioso (talk) — Iqr.png, CC BY-SA 3.0,

https ://commons.wikimedia.org/w/index.php ?curid=14702157

Quantiles et fonction de répartition

Si F bijective (ici : continue strictement croissante) :
Le quantile d'ordre $\alpha \in [0,1]$ est : $q_\alpha = F^{-1}(\alpha)$
q_α vérifie : $P(X \leqslant q_\alpha) = \alpha$.

Médiane : $q_{0,5}$ Quartiles : $q_{0,25}, q_{0,5}, q_{0,75}$



Pour la loi normale $\mathcal{N}(0,1)$, $q_{0,95} = 1,645$, et $q_{0,975} = 1,96$

Loi normale (ou loi de Gauss)

Une v.a. X suit une **loi normale** de moyenne μ et de variance σ^2 si sa densité est :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

On note : $X \sim \mathcal{N}(\mu, \sigma^2)$

$\mathcal{N}(0,1)$ est la **loi normale centrée réduite**.

Si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors $Z = (X - \mu)/\sigma \sim \mathcal{N}(0,1)$
On dit qu'on a centré et réduit X.
Z est appelé Z-score (plutôt avec estimateurs empiriques de μ et σ)

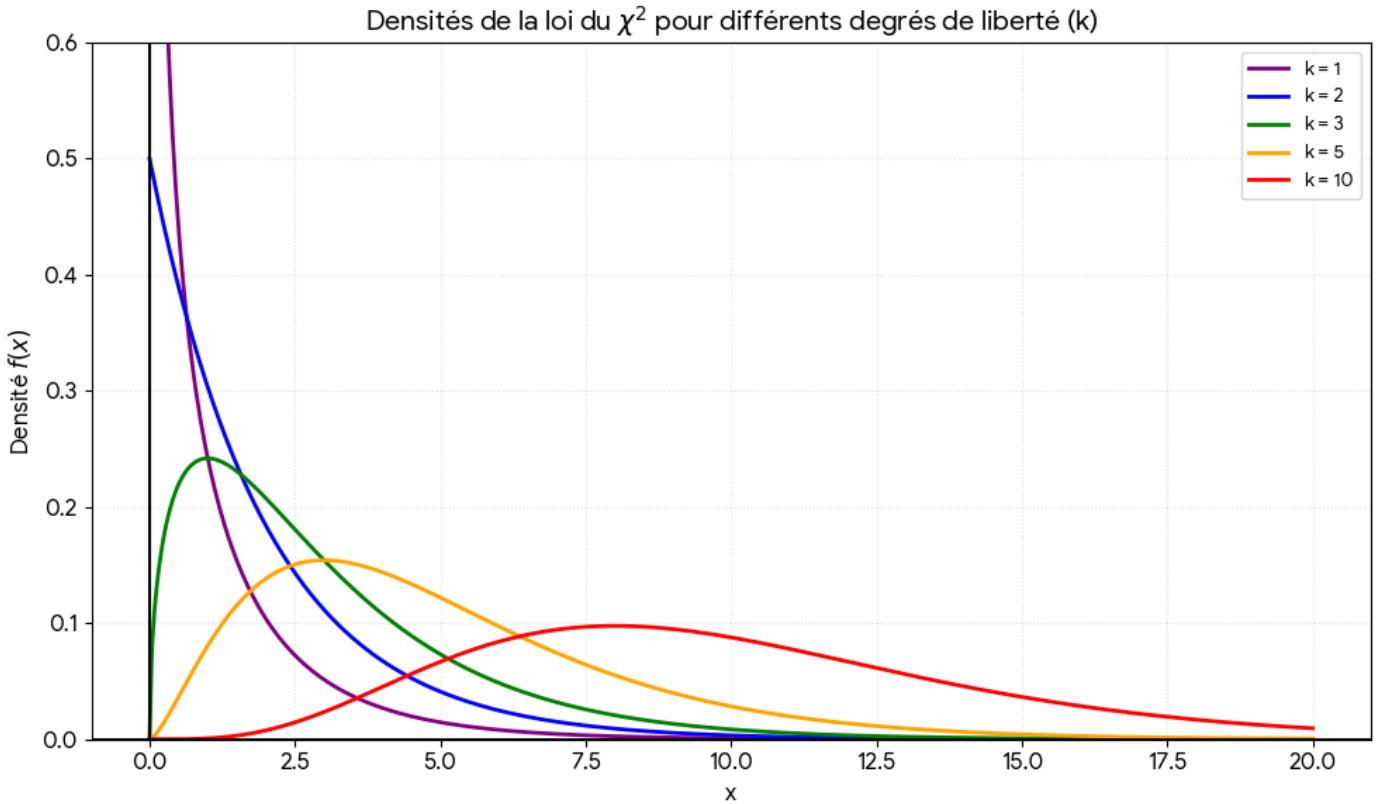
Pourquoi la loi normale est-elle si importante ?
$\rightarrow$ théorème central limite

Loi du χ^2

Soient $X_1, X_2, \ldots, X_k$ k v.a. indépendantes identiquement distribuées selon la loi $\mathcal{N}(0,1)$.

Par définition, $X = \sum_{i=1}^{k} X_i^2$ suit la **loi du χ^2 à k degrés de liberté**

On note $X \sim \chi^2_k$



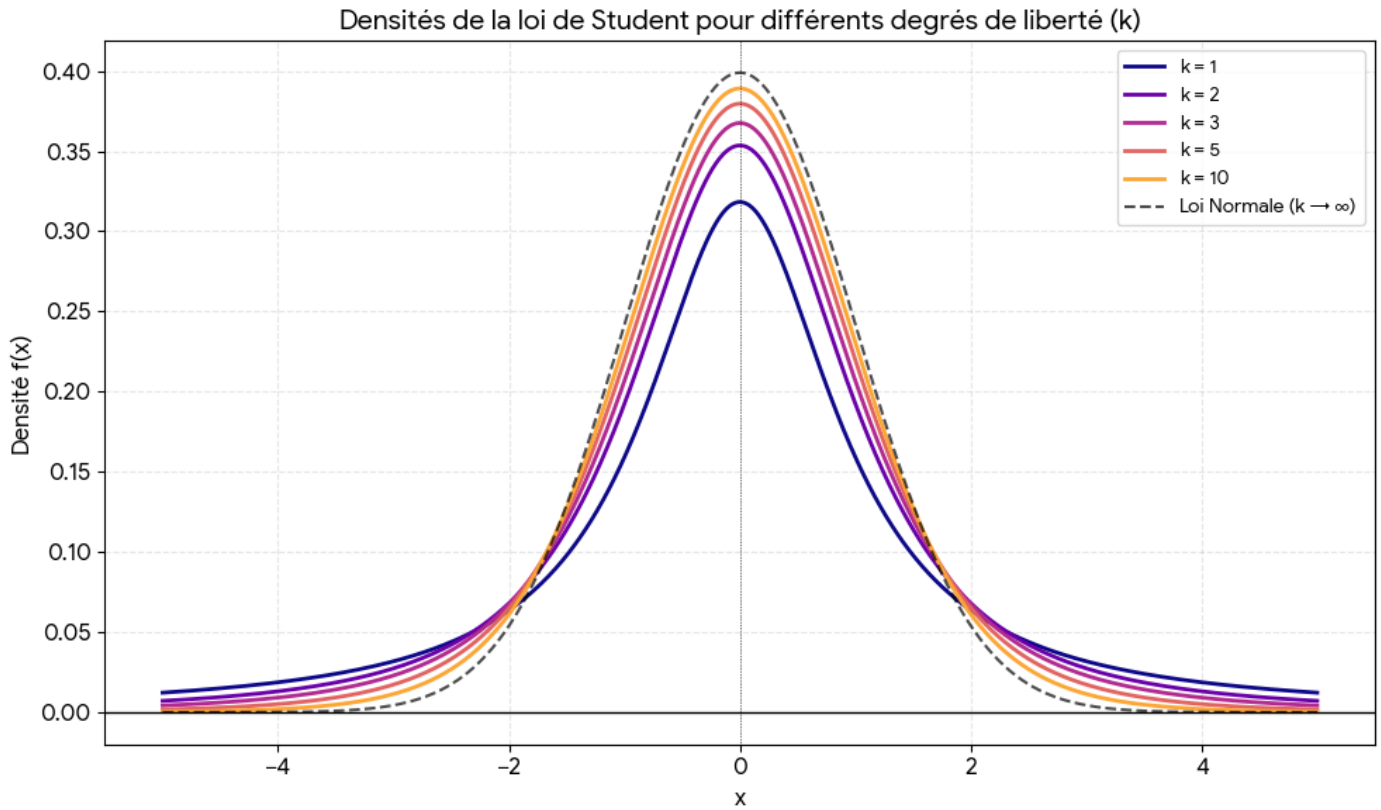
Propriétés : $E(X) = k$, $\mathrm{Var}(X) = 2k$

Loi de Student

Soient $Z \sim \mathcal{N}(0,1)$ et $U \sim \chi^2_k$ pour $k \geqslant 1$, indépendantes

Par définition, $T = \dfrac{Z}{\sqrt{U/k}}$ suit la **loi de Student à k d.d.l.**

On note $T \sim T_k$



Propriétés : si $k > 1, E(T) = 0$, si $k > 2, \mathrm{Var}(T) = k/(k-2)$

Loi de la variance d'un échantillon gaussien

Soit $(X_1, \ldots, X_n)$ un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$
$\rightarrow$ on sait que $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$

Soit S^2 la variance empirique (non corrigée)
alors $nS^2/\sigma^2 \sim \chi^2_{n-1}$
(voir l'exercice 3 de la séance 2 de la première partie du cours)
Avec S^* variance empirique corrigée : $(n-1)S^{*2}/\sigma^2 \sim \chi^2_{n-1}$

Conséquence :

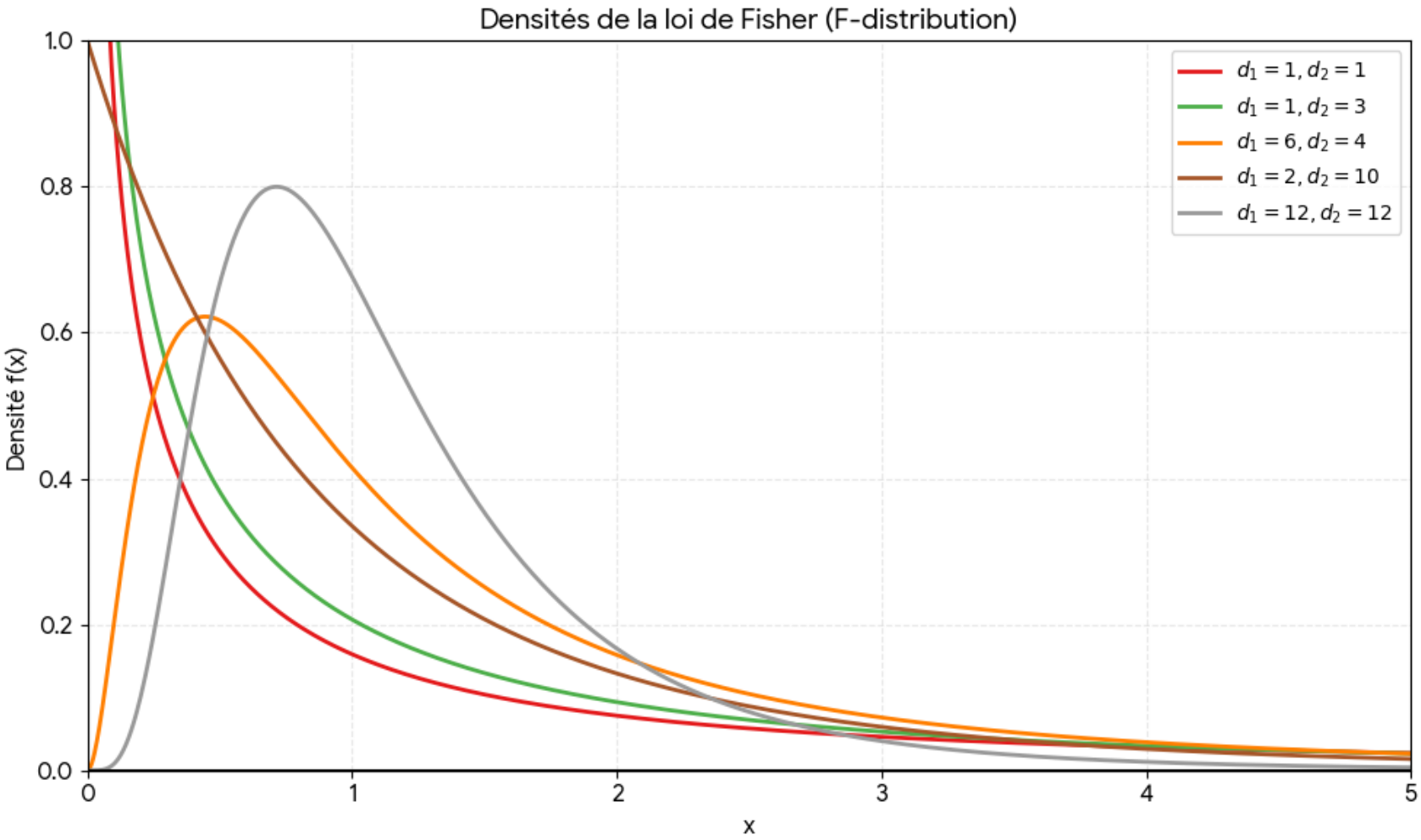
$$T = \frac{\overline{X} - \mu}{S^*/\sqrt{n}} \sim T_{n-1}$$

En effet, $T = \frac{Z}{\sqrt{U/(n-1)}}$ où $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ et $U = (n-1)S^{*2}/\sigma^2$

Loi de Fisher(-Snedecor)

Soient $U_1 \sim \chi^2_{d_1}$ et $U_2 \sim \chi^2_{d_2}$ indépendantes, alors

$F = \dfrac{U_1/d_1}{U_2/d_2}$ suit la **loi de Fisher-Snedecor à d_1 et d_2 d.d.l.**



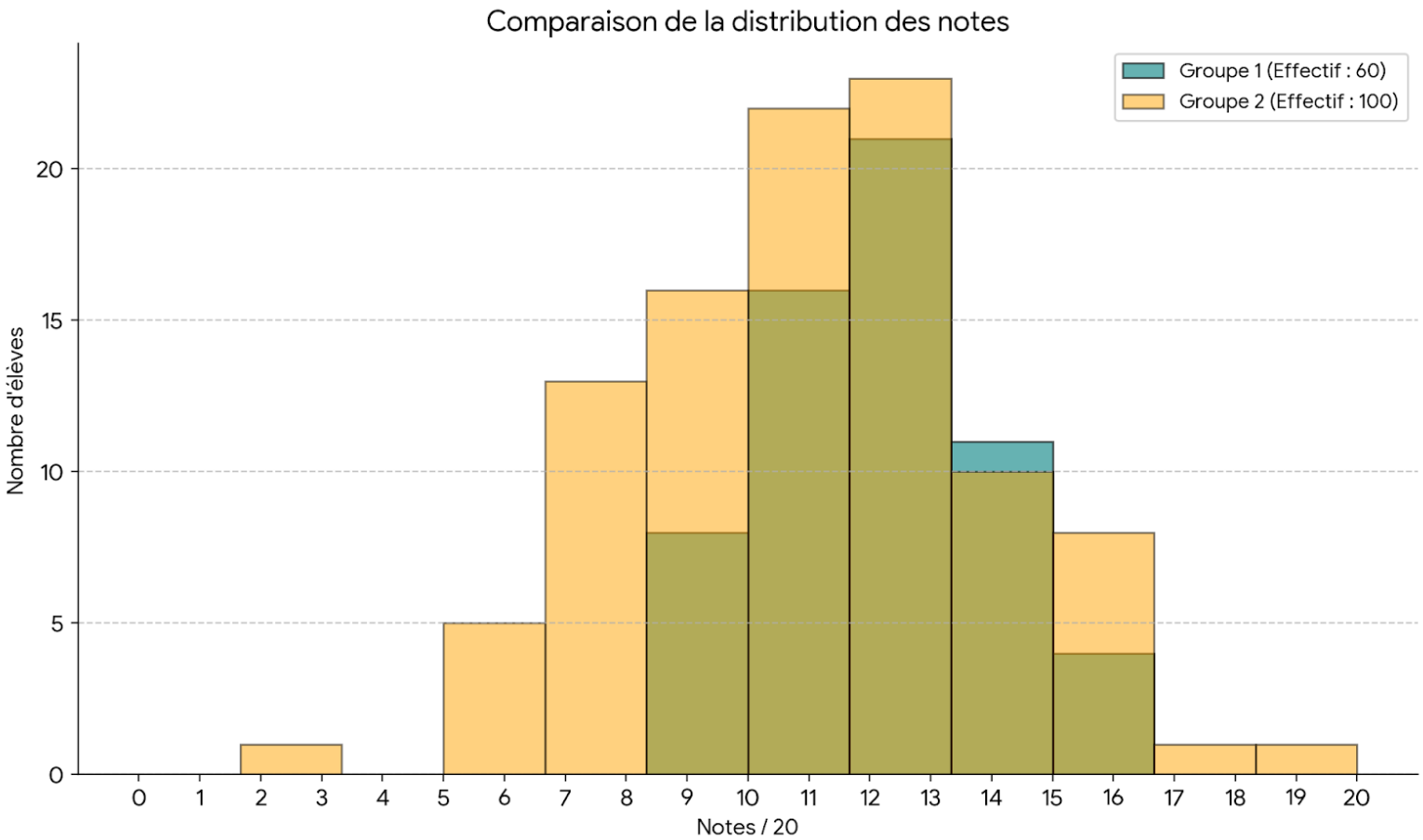
Plan

Cas de deux échantillons gaussiens

Voici les notes en maths de deux groupes parmi 160 élèves :

Groupe 1 : 60 élèves ayant suivi un programme renforcé en maths

Groupe 2 : 100 élèves n'ayant pas suivi ce programme



Question : Le programme a-t-il un effet ?

Test d'égalité des variances

Modélisation statistique :
Dans le groupe 1 : notes distribuées selon $\mathcal{N}(\mu_1, \sigma_1^2)$
Dans le groupe 2 : notes distribuées selon $\mathcal{N}(\mu_2, \sigma_2^2)$

Test d'égalité des variances : $\mathcal{H}_0 : \sigma_1^2 = \sigma_2^2$.
On calcule les variances empiriques corrigées :
$s_1^{*2} = 5,15 \quad s_2^{*2} = 5,71$

Voir cours : la statistique de test est $F = S_2^{*2}/S_1^{*2}$
rapport choisi pour être > 1, donc $\mathcal{H}_1 : \sigma_2^2/\sigma_1^2 > 1$

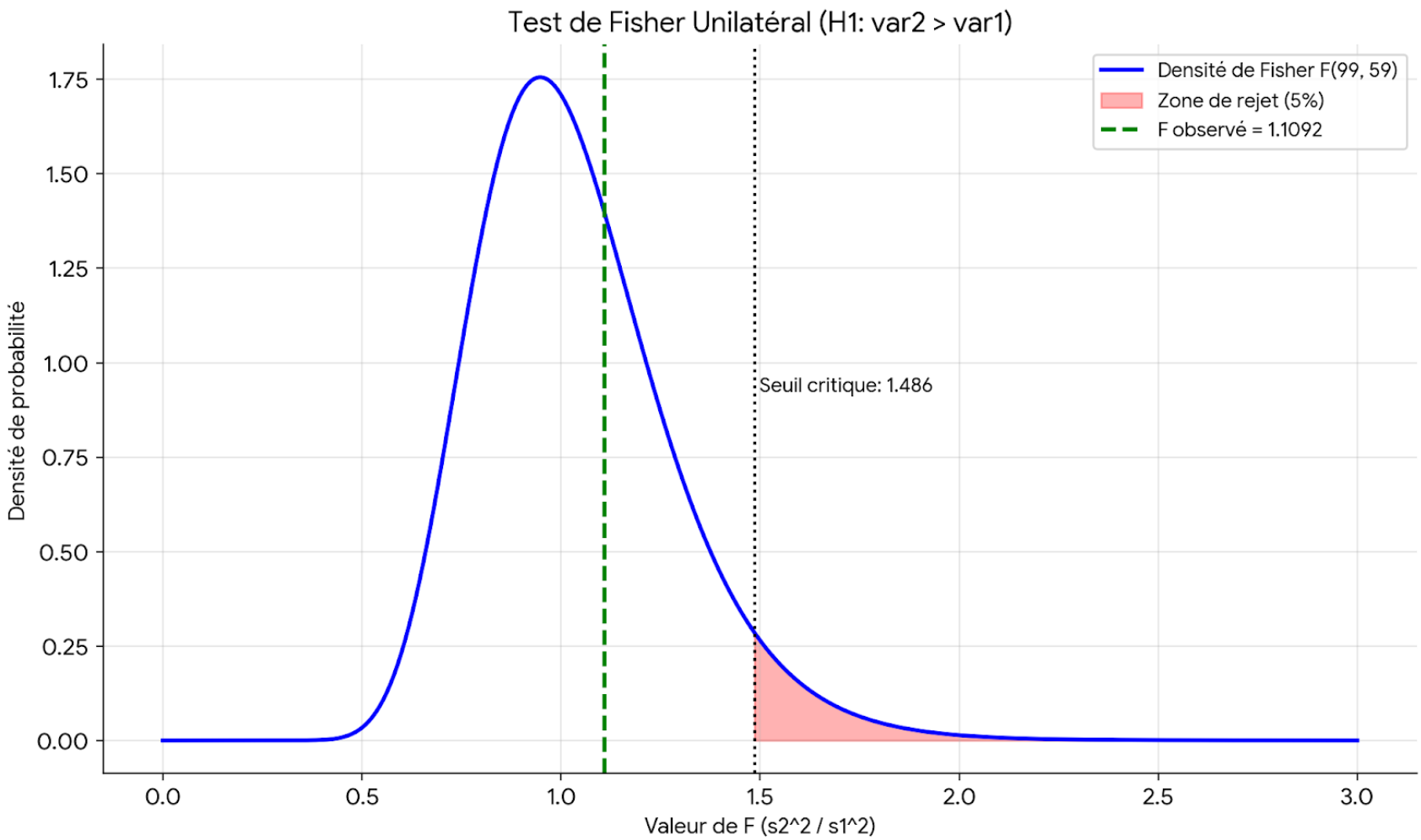
F suit une loi de Fisher-Snedecor à 99 et 59 degrés de liberté.

On calcule : $f_{\text{obs}} = s_2^{*2}/s_1^{*2} = 1,1092$

p-valeur pour un test unilatéral à droite : $P(F > 1,1092) = 0,3364$

Conclusion : on ne peut pas rejeter H_0 au risque de 5%.

Représentation graphique, comparaison au seuil critique



Seuil critique f_{crit} tel que $P(F > f_{\text{crit}}) = 0,05$
à comparer avec $f = 1,1092$

Test d'égalité des moyennes

Modélisation statistique :
Dans le groupe 1 : notes distribuées selon $\mathcal{N}(\mu_1, \sigma^2)$
Dans le groupe 2 : notes distribuées selon $\mathcal{N}(\mu_2, \sigma^2)$

Test d'égalité des moyennes : $\mathcal{H}_0 : \mu_1 = \mu_2$; $\mathcal{H}_1 : \mu_1 \neq \mu_2$
On calcule les moyennes empiriques :
$\bar{x}_1 = 11,605 \quad \bar{x}_2 = 11,458$
Voir la statistique de test dans le cours :
T suit une loi de Student à $n_1 + n_2 - 2$ d.d.l.

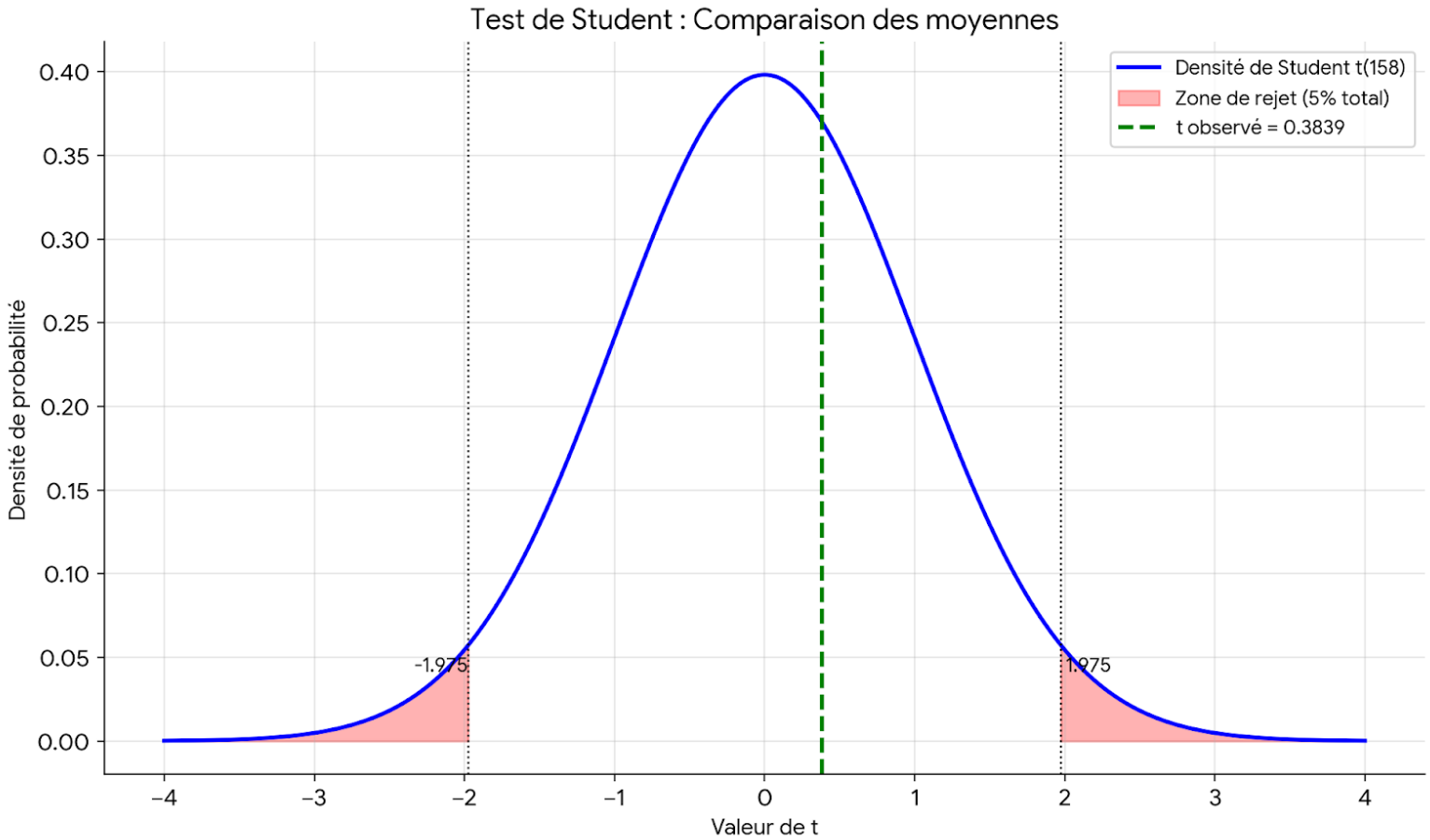
On calcule : $t_{\text{obs}} = 0,3839$

p-valeur pour un test bilatéral : $P(|T| > 0,3839) = 0,7016$

Conclusion : on ne peut pas rejeter H_0 au risque de 5%.

Attention : ne signifie pas que le programme renforcé ne sert pas !

Représentation graphique, comparaison au seuil critique



Remarque 1 : on peut se référer au quantile de la loi normale

Remarque 2 : si on veut tester $\mu_1 = \mu_2$ alors que $\sigma_1 \neq \sigma_2$?
$\rightarrow$ test t de Welsh.

Plan

Tests d'ajustement à une loi

Tests d'ajustement ou *tests d'adéquation* ou *tests de conformité*

Question : on observe la réalisation d'un n-échantillon. Cette observation est-elle cohérente avec l'hypothèse d'une distribution selon une loi donnée ?

Exemple (cours précédent) : des MTBF de 39,0 - 191,1 - 156,7 - 5.2 - 157.7 - 47.2 heures sont-ils cohérents avec une loi exponentielle de paramètre $\lambda = 1/100$?

Principe général : calculer une « distance » entre une distribution empirique (observée) et une distribution théorique (attendue).

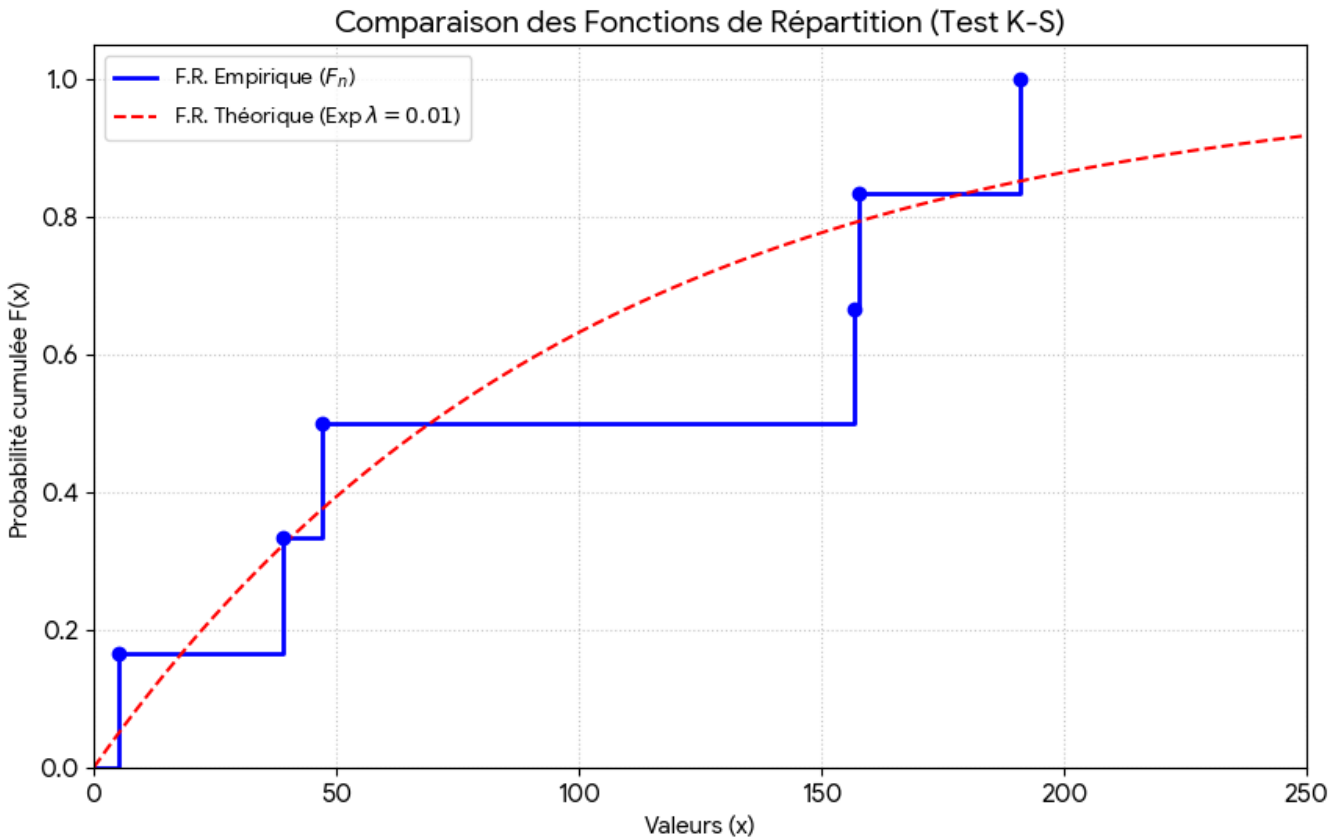
Fonctions de répartition empirique et théorique $P(X \leqslant x)$

Fonction de répartition empirique : $F_{\text{emp}}(x) = \frac{1}{n} \# \{i,\, X_i \leqslant x\}$

$\rightarrow$ réalisation $F_n(x) = \frac{1}{n} \# \{i,\, x_i \leqslant x\}$ (ici : $n = 6$)

Fonction de répartition théorique : $F_{\text{th}}(x) = 1 - \exp(-\lambda x)$

avec $\lambda = 1/100$



Test de Kolmogorov-Smirnov

Distance de Kolmogorov-Smirnov D_n : plus grand écart (vertical) entre les deux courbes

Sous hypothèse $\mathcal{H}_0$ que le n-échantillon suit la loi supposée, $\sqrt{n}D_n$ suit asymptotiquement la loi de Kolmogorov-Smirnov

$\rightarrow$ test unilatéral à droite car rejeter $\mathcal{H}_0$ signifie : $D_n \ll$ grand $\gg$.

Ici : $D_{\text{obs}} = 0,2914$

Valeur critique pour $n = 6$:
poly : $D_{\text{crit}} = 0,554$ $(= 1,358/\sqrt{6})$ (loi asymptotique)
table : $D_{\text{crit}} = 0,519$ (correction car n petit)

Résultat : comme $D_{\text{obs}} < D_{\text{crit}}$, on ne peut pas rejeter $\mathcal{H}_0$.

Plan

Test pour des variables qualitatives

Exemple : existe-t-il un lien entre le genre musical diffusé dans un point de vente et la catégorie de produits achetés par les clients ?

Table de contingence :

Genre X / Produit Y	Vêtements	Accessoires	Gadgets	**Total**
Classique	$n_{11} = 50$	$n_{12} = 30$	$n_{13} = 20$	$n_{1.} = 100$
Pop Rock	$n_{21} = 20$	$n_{22} = 40$	$n_{23} = 40$	$n_{2.} = 100$
Total	$n_{.1} = 70$	$n_{.2} = 70$	$n_{.3} = 60$	$n = 200$

$\mathcal{H}_0$: le choix du produit est indépendant du type de musique.

Effectifs théoriques attendus sous $\mathcal{H}_0$

Estimation de la loi des observations :

$P(X = i \text{ et } Y = j) \simeq n_{ij}/N$

$P(X = i) = \sum_{j=1}^{3} P(X = i \text{ et } Y = j) \simeq \sum_{j=1}^{3} n_{ij}/N = n_{i.}/n$

$P(Y = j) = \sum_{i=1}^{2} P(X = i \text{ et } Y = j) \simeq \sum_{i=1}^{2} n_{ij}/N = n_{.j}/n$

Loi théorique sous $\mathcal{H}_0$:

$P(X = i \text{ et } Y = j \mid \mathcal{H}_0) = P(X = i)P(Y = j) \simeq n_{i.}n_{.j}/n^2$

Effectifs attendus : $m_{ij} = n\,P(X = i \text{ et } Y = j \mid \mathcal{H}_0) = n_{i.}n_{.j}/n$

On calcule :

Genre X / Produit Y	Vêtements	Accessoires	Gadgets
Classique	$m_{11} = 35$	$n_{12} = 35$	$n_{13} = 30$
Pop Rock	$m_{21} = 35$	$n_{22} = 35$	$n_{23} = 30$

Distance du χ^2

$$d^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

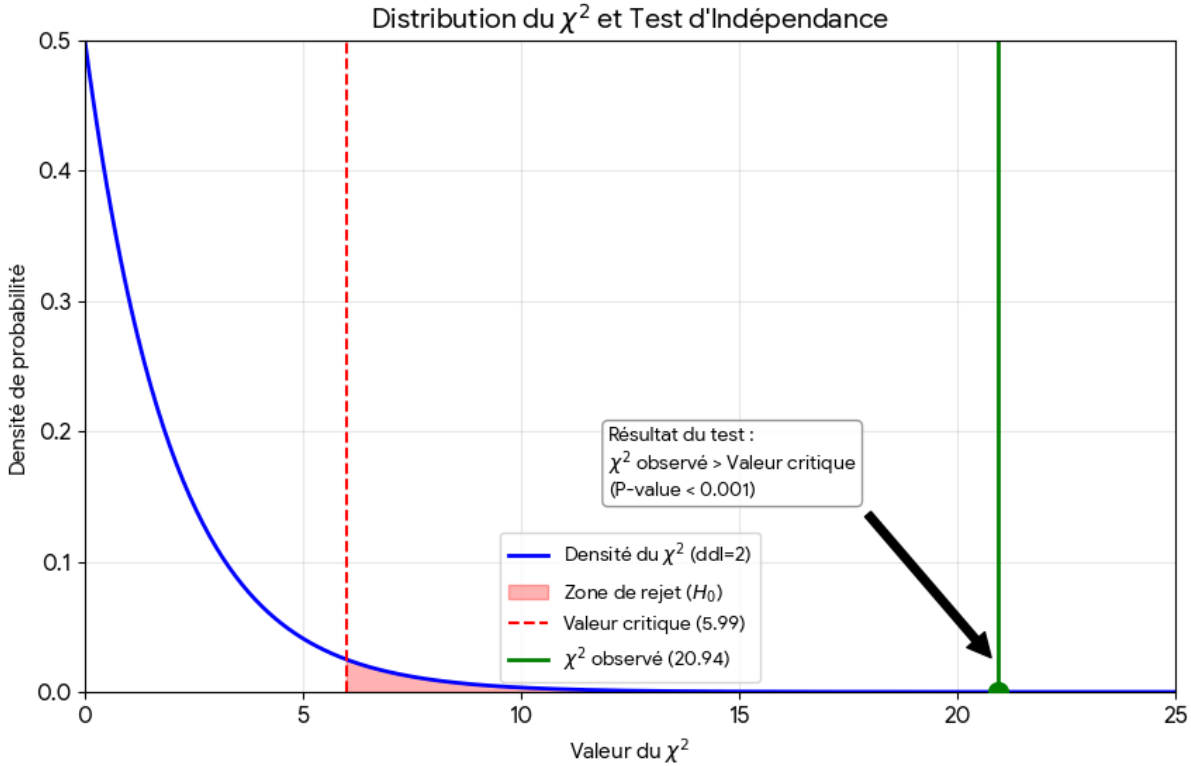
est (pour $n \ll$ grand $\gg$) réalisation d'une v.a. $\sim \chi^2_2$.

Critère de Cochran (1954) : $m_{ij} \geqslant 1$, et 80 % des classes $m_{ij} \geqslant 5 \ldots$

Calcul détaillé :

$$\begin{aligned} d^2 &= \frac{(50-35)^2}{35} + \frac{(30-35)^2}{35} + \frac{(20-30)^2}{30} \\ &\quad + \frac{(20-35)^2}{35} + \frac{(40-35)^2}{35} + \frac{(40-30)^2}{30} \\ d^2 &= 6,43 + 0,71 + 3,33 + 6,43 + 0,71 + 3,33 \\ d^2 &\approx 20,94 \end{aligned}$$

Test d'indépendance du χ^2



Conclusion : rejet de l'hypothèse $\mathcal{H}_0$

Attention : le test prouve une *association*, pas forcément une *causalité*. Une *variable de confusion* peut entraîner l'association : peut-être que la musique classique est diffusée le matin, moment où les clients achètent naturellement plus de vêtements.

Prochaine séance : étude de l'influence de différents facteurs.

Plan

Conclusion

Aujourd'hui :

1. Tests de comparaison de deux échantillons
   - variables gaussiennes : amphi + poly + ex 1
   - variables qualitatives : poly (test d'homogénéité du χ^2) + ex. 4-5 (test des signes)
2. Tests d'ajustement à une loi
   - ajustement une loi fixée, test de Kolmogorov-Smirnov : amphi + poly + ex 2a
   - test d'ajustement du χ^2 pour variables qualitatives : poly + ex 2b
3. Tests d'indépendance
   - tests de corrélation : poly
   - test d'indépendance du χ^2 : amphi + poly + ex 3