Cartographie des Tests Statistiques

Théorie, Applications et Implémentations

Abdelli Farès

Aout 2025

Table des matières

Introduction

Ce mémoire a pour but de proposer une cartographie rigoureuse et exhaustive des tests statistiques utilisés en science des données, en finance et en modélisation. Chaque test est présenté avec un souci de précision théorique et une mise en œuvre pratique sur des données réelles lorsque cela est possible. Le formalisme mathématique est explicitement détaillé, et l'implémentation s'appuie sur des bibliothèques Python couramment utilisées telles que `scipy`, `statsmodels` et `pandas`.

Exploration des données

1. Jeu de données `Titanic`

Le jeu de données `Titanic`, accessible via la bibliothèque `seaborn`, contient des informations sur les passagers du célèbre paquebot. Il est souvent utilisé pour illustrer des méthodes statistiques et d'apprentissage automatique.

1.1 Aperçu des données

```
import seaborn as sns
df = sns.load_dataset("titanic")
print(df.head())
```

Ce dataset comprend les colonnes suivantes :

Variable	Description
`survived`	1 = survivant, 0 = décédé
`pclass`	Classe du billet (1 = 1ère, 2 = 2e, 3 = 3e classe)
`sex`	Sexe (homme ou femme)
`age`	Âge (en années)
`sibsp`	Nombre de frères/sœurs ou conjoint à bord
`parch`	Nombre de parents/enfants à bord
`fare`	Prix du billet (en livres sterling)
`embarked`	Port d'embarquement (C = Cherbourg, Q = Queenstown, S = Southampton)
`class`	Classe du billet (catégorisé)
`who`	Homme, femme ou enfant
`adult_male`	Booléen indiquant si le passager est un homme adulte
`deck`	Pont du navire (A, B, C, etc.)
`embark_town`	Ville d'embarquement (nom complet)
`alive`	Statut de survie (yes/no)
`alone`	Booléen indiquant si le passager était seul

1.2 Statistiques descriptives

On présente ci-dessous un résumé statistique des variables numériques :

```
print(df.describe())
```

1.3 Analyse catégorielle

- **Taux de survie** : environ 38% des passagers ont survécu.
- **Répartition par classe** :
  - 1ère classe : $\sim$24%

Variable	Moyenne	Écart-type	Min	Max
age	29.70	14.52	0.42	80.0
fare	32.20	49.69	0.00	512.33
sibsp	0.52	1.10	0	8
parch	0.38	0.81	0	6

Table 2 – Statistiques descriptives des variables numériques

  - 2e classe : ∼21%
  - 3e classe : ∼55%
- **Répartition par sexe** :
  - Hommes : ∼64%
  - Femmes : ∼36%
- **Ports d'embarquement** :
  - Southampton : majoritaire
  - Cherbourg et Queenstown : minoritaires

1.4 Données manquantes

Certaines variables contiennent des valeurs manquantes :

- `age` : $\sim$ 20% de valeurs manquantes
- `deck` : très partiellement renseignée
- `embark_town` : quelques valeurs manquantes

Ces éléments doivent être pris en compte avant toute analyse statistique plus poussée (tests d'hypothèse, modélisation, etc.).

1 Tests de Normalité

1.1 Test de Shapiro-Wilk

But du test

Le test de Shapiro-Wilk est un test de normalité dont l'objectif est de déterminer si un échantillon de données suit une distribution normale. Il est particulièrement recommandé pour les petits échantillons (inférieurs à 50) mais reste valable jusqu'à 2000 observations.

Hypothèses

— H_0 : L'échantillon suit une distribution normale.
— H_1 : L'échantillon ne suit pas une distribution normale.

Formalisme mathématique

Soit $X = (x_1, \ldots, x_n)$ un échantillon de taille n, et $x_{(i)}$ les observations triées dans l'ordre croissant. Le test repose sur la statistique :

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

où $\bar{x}$ est la moyenne empirique et les coefficients a_i sont calculés à partir des moments d'une distribution normale standard. Plus précisément, si m est le vecteur des espérances des ordres statistiques d'une loi normale standard multivariée et V sa matrice de covariance, alors :

$$a = \frac{m^T V^{-1}}{\sqrt{m^T V^{-1} V^{-1} m}}$$

Statistique de test du test de Shapiro-Wilk

Le test de Shapiro-Wilk a pour objectif de tester l'hypothèse nulle selon laquelle un échantillon $X = (x_1, \ldots, x_n)$ est issu d'une population suivant une loi normale. Il est particulièrement adapté pour les petits et moyens échantillons ($n \leq 2000$).

La statistique du test est construite à partir des ordres statistiques de l'échantillon trié, notés $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, et de la moyenne empirique $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

La statistique de test W est définie par :

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

où les coefficients $(a_1, \ldots, a_n)$ dépendent uniquement de n et sont calculés à partir des moments d'une loi normale standard.

Plus précisément, si $m = (m_1, \ldots, m_n)^T$ est le vecteur des espérances des ordres statistiques d'un échantillon gaussien standard ($\mathcal{N}(0,1)$) de taille n, et V la matrice de covariance associée, alors les coefficients $a = (a_1, \ldots, a_n)^T$ sont donnés par :

$$a = \frac{m^T V^{-1}}{\sqrt{m^T V^{-1} V^{-1} m}}.$$

Les constantes a_i sont donc calculées une fois pour toutes pour chaque taille d'échantillon n, et tabulées ou pré-calculées dans les implémentations numériques (par exemple dans la fonction `shapiro()` de `scipy`).

La statistique W mesure la concordance entre la distribution de l'échantillon et celle attendue sous normalité : plus W est proche de 1, plus l'échantillon semble issu d'une loi normale. À l'inverse, des valeurs basses de W indiquent une déviation significative par rapport à la normalité.

La distribution exacte de W sous l'hypothèse nulle n'est pas triviale ; la p-valeur est généralement obtenue par simulations ou approximations numériques.

Intuition

Le numérateur du test mesure la proximité entre les données triées et leurs valeurs attendues sous une loi normale. Le dénominateur mesure la dispersion totale. Si les données sont normales, cette proximité sera élevée, donc W sera proche de 1. Des valeurs faibles de W indiquent une déviation significative par rapport à la normalité.

Implémentation Python

Afin d'évaluer la normalité de la variable `Age` dans le jeu de données `Titanic`, on applique le test de Shapiro-Wilk à l'aide de la fonction `shapiro()` de la bibliothèque `scipy.stats`. Il est ici appliqué sur les 60 premières valeurs non manquantes de la variable Age.

Nous complétons cette analyse par une visualisation graphique : un histogramme avec estimation de densité et un Q-Q plot permettant de juger visuellement de l'ajustement à la loi normale.

Le code Python correspondant est donné ci-dessous :

```python
# Importation des bibliothèques
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import shapiro

# Charger les données Titanic
df = pd.read_csv("titanic.csv")

# Supprimer les valeurs manquantes dans 'Age'et garder les 60 premières valeurs
age_data = df['Age'].dropna()[:60]

# Test de Shapiro-Wilk
statistic, p_value = shapiro(age_data)

print("Statistique de test W :", statistic)
print("P-valeur :", p_value)

# Décision
alpha = 0.05
if p_value < alpha:
    print("On rejette H0 : les données ne suivent pas une loi normale.")
else:
    print("On ne rejette pas H0 : les données peuvent être considérées comme normales.")

# Visualisation : histogramme et Q-Q plot
sns.histplot(age_data, kde=True)
plt.title("Distribution de la variable Age")
plt.show()

import scipy.stats as stats
stats.probplot(age_data, dist="norm", plot=plt)
plt.title("Q-Q Plot de Age")
plt.show()
```

Figure 1 – Code Shapiro

```
Statistique de test W : 0.9653247518485989
P-valeur : 0.08593139949277726
On ne rejette pas H0 : les données peuvent être considérées comme normales.
```

Figure 2 – Output Shapiro

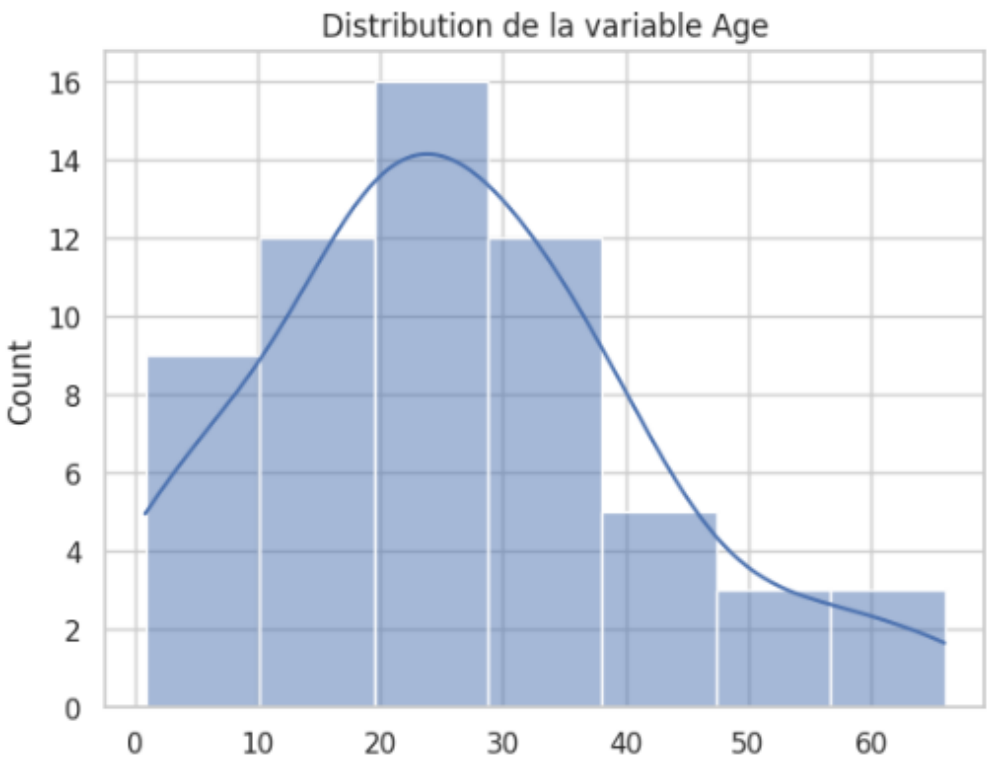

Figure 3 – KDE Shapiro

1.2 Test de Kolmogorov–Smirnov pour la normalité

Objectif :

Le test de Kolmogorov–Smirnov permet de tester si un échantillon i.i.d. $X_1, \ldots, X_n$ provient d'une loi normale standard $\mathcal{N}(0,1)$. On considère donc les hypothèses :

$$H_0 : F(x) = \Phi(x) \quad \text{vs} \quad H_1 : F(x) \neq \Phi(x)$$

où F est la fonction de répartition de la variable aléatoire sous-jacente, et Φ celle de la loi $\mathcal{N}(0,1)$.

Statistique de test :

On définit la fonction de répartition empirique de l'échantillon :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \leq x\}}$$

La statistique du test est alors :

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - \Phi(x)|$$

Elle mesure l'écart maximal entre la fonction de répartition empirique et la fonction de répartition théorique.

Comportement asymptotique de la statistique sous H_0 :

Sous l'hypothèse nulle, les X_i sont i.i.d. de loi $\mathcal{N}(0,1)$. Le processus centré et normalisé :

$$\sqrt{n}(F_n(x) - \Phi(x))$$

converge en loi dans l'espace des fonctions càdlag vers un processus stochastique limite noté $B_0(t)$, où $t = \Phi(x) \in [0,1]$. Ce processus est appelé le *pont brownien*.

Définition du pont brownien

Le pont brownien $B_0(t)$, $t \in [0,1]$, est un processus gaussien défini par :

$$B_0(t) = W(t) - tW(1)$$

où $W(t)$ est un mouvement brownien standard. Ce processus satisfait $B_0(0) = B_0(1) = 0$ presque sûrement, d'où le nom de "pont".

Loi limite de la statistique

La convergence en loi suivante est alors valable sous H_0 :

$$\sqrt{n}D_n \xrightarrow[n\to\infty]{\mathcal{L}} \sup_{t\in[0,1]} |B_0(t)|$$

La variable aléatoire limite $\sup_{t\in[0,1]} |B_0(t)|$ suit une loi appelée *loi de Kolmogorov*, dont la fonction de répartition est :

$$K(\lambda) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2\lambda^2}, \quad \lambda > 0$$

Règle de décision

On fixe un seuil α (par exemple 5%). On rejette l'hypothèse nulle H_0 si :

$$\sqrt{n}D_n > K^{-1}(1-\alpha)$$

où K^{-1} désigne la fonction quantile de la loi limite. Alternativement, des tables de valeurs critiques de D_n pour divers n sont utilisées.

Intuition derrière le test

La fonction de répartition empirique est une estimation naturelle et simple de la distribution sous-jacente des données observées. En comparant cette estimation à la fonction théorique, on teste la concordance entre les données observées et la loi hypothétique. La prise du maximum de la différence assure que même une petite zone de discordance importante peut suffire à rejeter l'hypothèse nulle, ce qui rend le test sensible aux écarts locaux entre distributions.

—

Domaines d'application

Le test de Kolmogorov-Smirnov est largement utilisé dans différents domaines comme :

— La finance, pour vérifier la conformité des rendements financiers à des modèles de distribution.

— Le contrôle qualité et la recherche scientifique, pour valider l'adéquation des modèles statistiques.

Il est particulièrement apprécié pour sa flexibilité puisqu'il ne nécessite pas que la variance ou d'autres paramètres soient connus, contrairement aux tests paramétriques.

Implémentation Python

Pour évaluer la normalité des données d'âge issues du jeu de données *Titanic*, nous appliquons le test de Kolmogorov-Smirnov. Les données sont d'abord centrées et réduites, puis comparées à une distribution normale standard $\mathcal{N}(0,1)$.

```python
from scipy.stats import kstest, norm

# Charger les données Titanic
df = pd.read_csv("titanic.csv")

df.head()
# Extraire les âges valides (non NaN)
age_data = df['Age'].dropna()[:250]

# Centrer et réduire les données (standardisation)
mean = age_data.mean()
std = age_data.std()
standardized_ages = (age_data - mean) / std

# Test K-S : comparer aux quantiles d'une N(0,1)
stat, p_value = kstest(standardized_ages, 'norm')

print(f"Statistique de test D : {stat:.4f}")
print(f"P-valeur : {p_value:.4e}")

# Décision
alpha = 0.05
if p_value < alpha:
    print("On rejette H0 : les données ne suivent pas une loi normale.")
else:
    print("On ne rejette pas H0 : les données peuvent suivre une loi normale.")
```

Figure 4 – Code KS

```
Statistique de test D : 0.0845
P-valeur : 5.2933e-02
On ne rejette pas H0 : les données peuvent suivre une loi normale.
```

Figure 5 – Output KS

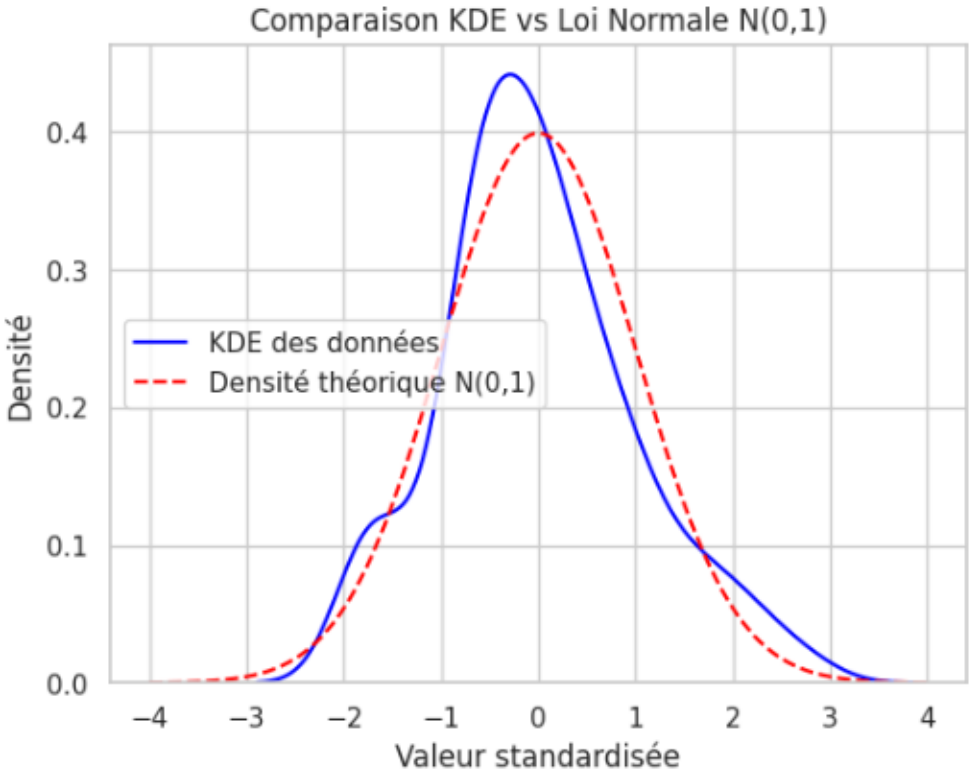

Figure 6 – KDE KS

2 Test de la moyenne avec variance connue (Test Z)

1. Objectif du test

Ce test permet de vérifier si la moyenne d'une population gaussienne μ est égale à une valeur théorique μ_0, dans le cas où la variance σ^2 de la population est connue.

Il est souvent utilisé en contrôle qualité, en ingénierie, ou dans des contextes où la variance est estimée de manière fiable par des études précédentes ou des propriétés physiques bien connues.

2. Hypothèses du test

— $\mathcal{H}_0 : \mu = \mu_0$
— $\mathcal{H}_1 : \mu \neq \mu_0$ (test bilatéral), ou $\mu > \mu_0$, ou $\mu < \mu_0$ (test unilatéral)

3. Conditions d'application

— La variable aléatoire X suit une loi normale : $X \sim \mathcal{N}(\mu, \sigma^2)$
— La variance σ^2 est connue
— Les observations sont indépendantes

4. Statistique de test

Soit un échantillon $x_1, x_2, \ldots, x_n$, on note :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

La statistique de test est :

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Sous $\mathcal{H}_0$, cette statistique suit une **loi normale centrée réduite** $\mathcal{N}(0, 1)$.

Convergence asymptotique de la statistique du test Z

Considérons un échantillon de taille n issu d'une population de moyenne μ et d'écart-type σ connu. Soit $\bar{X}$ la moyenne empirique :

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

Par le théorème central limite (TCL), lorsque n est suffisamment grand, la variable aléatoire $\bar{X}$ suit approximativement une loi normale :

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

où :
— μ est la moyenne réelle de la population,
— σ est l'écart-type connu,
— $\xrightarrow{d}$ indique une convergence en distribution.
La statistique de test définie par :

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

suit donc asymptotiquement une loi normale standard $\mathcal{N}(0,1)$ sous l'hypothèse nulle $H_0 : \mu = \mu_0$.

Cette convergence permet d'utiliser la loi normale pour déterminer les régions critiques et les p-valeurs, même si la distribution initiale de X_i n'est pas normale, à condition que n soit suffisamment grand.

5. Règle de décision

— Pour un test bilatéral au niveau α : on rejette $\mathcal{H}_0$ si $|Z| > z_{1-\alpha/2}$
— Pour un test unilatéral à droite : on rejette $\mathcal{H}_0$ si $Z > z_{1-\alpha}$
— Pour un test unilatéral à gauche : on rejette $\mathcal{H}_0$ si $Z < -z_{1-\alpha}$
où $z_{1-\alpha}$ est le quantile de la loi normale centrée réduite.

6. Intervalle de confiance associé au test Z

Le test Z permet également de construire un intervalle de confiance pour la moyenne μ lorsque l'écart-type σ de la population est connu.

L'intervalle de confiance au niveau de confiance $1 - \alpha$ est donné par :

$$\left[\bar{X} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

où :
— $\bar{X}$ est la moyenne de l'échantillon,
— $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale standard $\mathcal{N}(0,1)$,
— σ est l'écart-type connu de la population,
— n est la taille de l'échantillon.
Par exemple, pour un niveau de confiance de 95%, on a $z_{0.975} \approx 1.96$.

Cet intervalle contient la valeur réelle de la moyenne avec une probabilité de $1 - \alpha$ sur un grand nombre d'échantillons répétés.

Exemple : test sur un échantillon gaussien avec variance connue

Nous illustrons ici l'application du test z de la moyenne lorsque la variance de la population est supposée connue. La décision est prise selon la règle classique de rejet basée sur le dépassement d'un seuil critique.

```python
import numpy as np
from scipy.stats import norm

# Paramètres
mu_0 = 100
sigma = 15
alpha = 0.05
n = 50

# Génération des données
np.random.seed(0)
data = np.random.normal(loc=mu_0, scale=sigma, size=n)

# Moyenne de l'échantillon
mean_sample = np.mean(data)

# Seuil critique pour un test bilatéral
z_critique = norm.ppf(1 - alpha/2)
borne = z_critique * sigma / np.sqrt(n)

# Calcul de la statistique
ecart_obs = abs(mean_sample - mu_0)

# Affichage
print(f"Moyenne observée : {mean_sample:.2f}")
print(f"Écart observé : {ecart_obs:.3f}")
print(f"Borne critique : {borne:.3f}")

# Décision
if ecart_obs > borne:
    print("On rejette H0 : la moyenne diffère significativement de
       100.")
else:
    print("On ne rejette pas H0 : la moyenne n'est pas
       significativement différente de 100.")
```

Sortie typique :

```
Moyenne observée : 102.12
Écart observé : 2.118
Borne critique : 4.160
→ On ne rejette pas H0 : la moyenne n'est pas significativement différente de
   100.
```

3 Test t de Student pour la moyenne (variance inconnue)

3.1 Objectif du test

Ce test permet de déterminer si la moyenne μ d'une population normale est égale à une valeur fixée μ_0, lorsque la variance σ^2 de la population est inconnue. Il s'agit d'un test paramétrique, adapté à de petits échantillons ($n < 30$) où l'estimation de la variance introduit de l'incertitude.

3.2 Hypothèses

— $H_0 : \mu = \mu_0$ (la moyenne hypothétique est correcte)
— $H_1 : \mu \neq \mu_0$ (la moyenne diffère de μ_0) — cas bilatéral.

3.3 Conditions d'application

— Les observations $X_1, \ldots, X_n$ sont i.i.d. (indépendantes et identiquement distribuées).
— La variable X suit une distribution normale.
— La variance σ^2 est inconnue.

3.4 Statistique de test

Lorsque la variance est inconnue, on l'estime à partir de l'échantillon. On définit :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

La statistique de test est alors :

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

3.5 Loi de la statistique sous H_0

Sous l'hypothèse nulle H_0, si $X_i \sim \mathcal{N}(\mu, \sigma^2)$, alors la statistique T suit une loi de Student à $n-1$ degrés de liberté :

$$T \sim t_{n-1}$$

Cela est dû au fait que l'estimateur S^2 suit une loi du χ^2 et qu'il est indépendant de $\bar{X}$.

3.6 Décision

On choisit un seuil de signification α (par exemple, 5%) et on calcule la valeur critique $t_{\alpha/2,n-1}$ à partir de la table de la loi de Student.
— Si $|T| > t_{\alpha/2,n-1}$, on **rejette** H_0.
— Sinon, on **ne rejette pas** H_0.

3.7 Remarque sur le TCL

Lorsque n est grand ($n \geq 30$), la loi de Student converge vers la loi normale standard, et on peut alors approximer T par une loi normale $\mathcal{N}(0,1)$.

3.8 Convergence de la statistique de test du t-test (variance inconnue)

Rappel du cadre

Soit $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, où :
— μ est la moyenne inconnue à tester ;
— σ^2 est la variance inconnue.
On considère l'estimateur empirique de la moyenne :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

et l'estimateur empirique non biaisé de la variance :

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Distribution de S^2 sous H_0

On peut montrer que :

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Justification : Si $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, alors les variables centrées réduites $Z_i = \frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0,1)$. En exprimant la variance empirique via ces Z_i, on obtient :

$$(n-1) \cdot \frac{S^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2.$$

Ce terme est une somme de carrés de $n-1$ variables indépendantes suivant une loi normale centrée réduite (car une seule contrainte lie les X_i via $\bar{X}$), donc :

$$(n-1) \cdot \frac{S^2}{\sigma^2} \sim \chi^2(n-1).$$

Indépendance de $\bar{X}$ et S^2

Un résultat des lois normales est que, lorsque les X_i sont i.i.d. normales :

$$\bar{X} \text{ et } S^2 \text{ sont indépendants.}$$

Ce résultat ne vaut que pour des données gaussiennes.

Statistique de test

La statistique du test de Student est définie par :

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

On souhaite déterminer la loi de T sous H_0.

Lien avec la loi de Student

On peut écrire :

$$T = \underbrace{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}_{\sim \mathcal{N}(0,1)} \cdot \underbrace{\frac{\sigma}{S}}_{= \left(\frac{\chi^2(n-1)}{n-1}\right)^{-1/2}}.$$

Posons :

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1), \quad V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1), \quad \text{avec } Z \perp V.$$

Alors on a :

$$T = \frac{Z}{\sqrt{V/(n-1)}} \sim \mathcal{T}_{n-1},$$

où $\mathcal{T}_{n-1}$ désigne la loi de Student à $n-1$ degrés de liberté.

3.9 Conclusion

La statistique T suit une loi de Student, car elle est construite comme le ratio :

$$\frac{\text{normale centrée réduite}}{\sqrt{\text{chi² indépendante / ddl}}},$$

ce qui **définit** la loi $\mathcal{T}_{n-1}$.

Ce résultat est valable uniquement sous l'hypothèse que les données sont issues d'une population normale.

—

3.10 Intervalle de confiance pour la moyenne (variance inconnue)

Lorsque la variance de la population est inconnue, on peut construire un intervalle de confiance pour la moyenne μ en utilisant la loi de Student.

L'intervalle de confiance au niveau de confiance $1-\alpha$ est donné par :

$$\left[\bar{X} - t_{1-\frac{\alpha}{2},n-1} \cdot \frac{S}{\sqrt{n}},\ \bar{X} + t_{1-\frac{\alpha}{2},n-1} \cdot \frac{S}{\sqrt{n}}\right]$$

où :

— $\bar{X}$ est la moyenne empirique,
— S est l'écart-type empirique de l'échantillon,
— $t_{1-\frac{\alpha}{2},n-1}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Student à $n-1$ degrés de liberté,
— n est la taille de l'échantillon.

Cet intervalle est exact si les données sont issues d'une population normale, et asymptotiquement correct lorsque n est grand, même en dehors du cadre normal.

3.11 Implémentation python

Afin de tester si la moyenne d'âge des passagers du *Titanic* diffère significativement d'une valeur hypothétique (ici 30 ans), nous réalisons le test sur un échantillon de 250 individus.

```python
import pandas as pd
from scipy import stats

# Charger le dataset Titanic
df = pd.read_csv("titanic.csv")

# Supprimer les lignes avec valeurs manquantes dans "age"
age_data = df["Age"].dropna()[:250]

# Paramètre d'hypothèse nulle
mu_0 = 30  # valeur hypothétique de la moyenne

# Calcul de la statistique de test
t_statistic, p_value = stats.ttest_1samp(age_data, popmean=mu_0)

# Affichage des résultats
print(f"Statistique t : {t_statistic:.4f}")
print(f"p-value : {p_value:.4f}")

# Interprétation à alpha = 0.05
alpha = 0.05
if p_value < alpha:
    print("On rejette H0 : la moyenne des âges est significativement différente de 30.")
else:
    print("On ne rejette pas H0 : la moyenne des âges n'est pas significativement différente de 30.")
```

Figure 7 – Code test T de Student

```
Statistique t : -1.4075
p-value : 0.1605
On ne rejette pas H0 : la moyenne des âges n'est pas significativement différente de 30.
```

Figure 8 – Output du test

4 Test t de Student à deux échantillons indépendants

Objectif du test

Le test t de Student à deux échantillons indépendants a pour but de comparer les moyennes de deux populations supposées gaussiennes, lorsque les variances des deux populations sont inconnues. Ce test est l'analogue bilatéral du test de la moyenne pour un seul échantillon, appliqué ici à deux groupes distincts.

Cadre et hypothèses

Soient deux échantillons indépendants :
— $X^{(1)} = (X_1^{(1)}, X_2^{(1)}, \ldots, X_{n_1}^{(1)})$ de taille n_1, issu d'une population de moyenne μ_1 et de variance σ_1^2 ;
— $X^{(2)} = (X_1^{(2)}, X_2^{(2)}, \ldots, X_{n_2}^{(2)})$ de taille n_2, issu d'une population de moyenne μ_2 et de variance σ_2^2 ;
On suppose que :

$$X_i^{(1)} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2), \quad X_i^{(2)} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2), \quad X^{(1)} \perp X^{(2)}$$

On souhaite tester l'hypothèse :

$$H_0 : \mu_1 = \mu_2 \quad \text{contre} \quad H_1 : \mu_1 \neq \mu_2$$

Les variances σ_1^2 et σ_2^2 étant inconnues, deux cas sont à distinguer selon qu'elles soient supposées égales ou non.

4.1 Variances inconnues mais supposées égales

Estimations :

$$\bar{X}^{(1)} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i^{(1)}, \quad \bar{X}^{(2)} = \frac{1}{n_2} \sum_{j=1}^{n_2} X_j^{(2)}$$

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(X_i^{(1)} - \bar{X}^{(1)} \right)^2, \quad s_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} \left(X_j^{(2)} - \bar{X}^{(2)} \right)^2$$

On estime la variance commune :

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Statistique de test :

$$T = \frac{(\bar{X}^{(1)} - \bar{X}^{(2)}) - (\mu_1 - \mu_2)}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Sous H_0, cette statistique suit une loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Convergence de la statistique de test

La statistique de test pour le cas des deux échantillons indépendants, à variances inconnues mais supposées égales, est donnée sous H_0 par :

$$T = \frac{\bar{X}^{(1)} - \bar{X}^{(2)}}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Nous allons étudier la convergence en loi de cette statistique lorsque $n_1 \to \infty$, $n_2 \to \infty$, et sous l'hypothèse nulle $H_0 : \mu_1 = \mu_2$.

On peut décomposer T comme le produit de deux termes :

$$T = \underbrace{\frac{\bar{X}^{(1)} - \bar{X}^{(2)}}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}_{\text{Suit } \mathcal{N}(0,1)} \cdot \underbrace{\frac{\sigma}{s_p}}_{\xrightarrow{\mathbb{P}} 1}$$

Comportement du premier terme Sous H_0, on a $\mu_1 = \mu_2$, donc $\bar{X}^{(1)} - \bar{X}^{(2)}$ est centré, et :

$$\bar{X}^{(1)} - \bar{X}^{(2)} \sim \mathcal{N}\left(0, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right) \Rightarrow \frac{\bar{X}^{(1)} - \bar{X}^{(2)}}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N}(0,1)$$

Comportement du dénominateur On rappelle que la variance commune est estimée par :

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

où

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i^{(1)} - \bar{X}^{(1)})^2 \quad \text{et} \quad s_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(X_j^{(2)} - \bar{X}^{(2)})^2$$

Or, sous l'hypothèse que chaque échantillon suit une loi normale $\mathcal{N}(\mu_i, \sigma^2)$, on sait que :

— $(n_1 - 1)\frac{s_1^2}{\sigma^2} \sim \chi^2_{n_1 - 1}$

— $(n_2 - 1)\frac{s_2^2}{\sigma^2} \sim \chi^2_{n_2 - 1}$ (voir annexe)

— Les deux statistiques sont indépendantes

Donc :

$$\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{\sigma^2} \sim \chi^2_{n_1 + n_2 - 2}$$

On a ainsi :

$$\frac{s_p^2}{\sigma^2} = \frac{\chi^2_{n_1+n_2-2}}{n_1 + n_2 - 2} \xrightarrow{\mathbb{P}} 1 \quad \Rightarrow \quad s_p \xrightarrow{\mathbb{P}} \sigma$$

Cette convergence utilise le fait que $\chi^2_k / k \xrightarrow{\mathbb{P}} 1$ lorsque $k \to \infty$ (loi des grands nombres).

Application du théorème de Slutsky. Pour de grand échantillons, puisque le numérateur de T converge en loi vers une normale centrée réduite, et que le dénominateur σ/s_p converge en probabilité vers 1, on peut appliquer le théorème de Slutsky :

$$T \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Cette convergence est uniquement valable asymptotiquement. En réalité, sous H_0, et pour tout n_1, n_2, la statistique de test suit exactement une loi de Student à $n_1 + n_2 - 2$ degrés de liberté :

$$T \sim \mathcal{T}_{n_1+n_2-2}$$

Conclusion La statistique de test T, qui suit exactement une loi de Student à $n_1 + n_2 - 2$ degrés de liberté sous H_0, converge asymptotiquement vers une loi normale standard lorsque les tailles d'échantillons deviennent grandes. Cela justifie l'utilisation d'une approximation normale pour le test t dans les grands échantillons, même lorsque la variance est inconnue.

Règle de décision :

$$\text{On rejette } H_0 \text{ au niveau } \alpha \text{ si } |T| > t_{1-\alpha/2,n_1+n_2-2}$$

4.2 Variances inconnues et supposées différentes (Test de Welch)

Cadre et hypothèses : Soient deux échantillons indépendants :

— $X_1^{(1)}, \ldots, X_{n_1}^{(1)} \overset{iid}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$;

— $X_1^{(2)}, \ldots, X_{n_2}^{(2)} \overset{iid}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$;

avec $X^{(1)} \perp X^{(2)}$.

On souhaite tester :

$$H_0 : \mu_1 = \mu_2 \quad \text{contre} \quad H_1 : \mu_1 \neq \mu_2.$$

Statistique de test :

$$T = \frac{\bar{X}^{(1)} - \bar{X}^{(2)} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Sous $H_0 : \mu_1 = \mu_2$, cette statistique devient :

$$T = \frac{\bar{X}^{(1)} - \bar{X}^{(2)}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Justification de la loi (approximative) de Student : Sous l'hypothèse de normalité des deux échantillons :

— $\bar{X}^{(1)} \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$, $\bar{X}^{(2)} \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$;

— $\frac{(n_1-1)s_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$, $\frac{(n_2-1)s_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$;

— $\bar{X}^{(1)}$, s_1^2 et $\bar{X}^{(2)}$, s_2^2 sont indépendants.

Ainsi, le numérateur est normal centré sous H_0, et le dénominateur est une racine de somme pondérée de deux χ^2 indépendantes, divisées par leurs degrés de liberté respectifs. Cette situation n'est pas exactement celle d'un χ^2 mais peut être approchée par une loi de Student à ν degrés de liberté, via l'approximation de Satterthwaite.

Degrés de liberté effectifs :

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

Conclusion : Sous H_0, et avec les hypothèses de normalité, la statistique T suit approximativement une loi de Student à ν degrés de liberté. Cette approximation devient exacte lorsque σ_1^2 et σ_2^2 sont connues, ou lorsque n_1 et n_2 deviennent grands (théorème central limite et approximation de Satterthwaite).

Convergence asymptotique de la statistique de test On considère la statistique de test suivante, sous l'hypothèse nulle $H_0 : \mu_1 = \mu_2$:

$$T = \frac{\bar{X}^{(1)} - \bar{X}^{(2)}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

1. Convergence du numérateur.
Sous H_0, on a :

$$\bar{X}^{(1)} - \bar{X}^{(2)} = \left(\bar{X}^{(1)} - \mu_1\right) - \left(\bar{X}^{(2)} - \mu_2\right)$$

Chaque moyenne empirique est une moyenne d'observations iid. D'après le **théorème central limite** (TCL), on a :

$$\bar{X}^{(1)} \xrightarrow{\mathcal{L}} \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \quad \bar{X}^{(2)} \xrightarrow{\mathcal{L}} \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

Comme les deux échantillons sont indépendants, la différence converge vers une normale de moyenne nulle et de variance égale à la somme des deux variances :

$$\bar{X}^{(1)} - \bar{X}^{(2)} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

2. Convergence du dénominateur.
On rappelle que :

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(X_i^{(1)} - \bar{X}^{(1)}\right)^2$$

Cet estimateur est sans biais de la variance σ_1^2, et on peut montrer que :

$$s_1^2 \xrightarrow{\mathbb{P}} \sigma_1^2, \quad s_2^2 \xrightarrow{\mathbb{P}} \sigma_2^2$$

Ceci découle de la **loi des grands nombres** appliquée à une suite de variables indépendantes et identiquement distribuées ayant une variance finie.

Ainsi, le dénominateur :

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \xrightarrow{\mathbb{P}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

3. Conclusion par le théorème de Slutsky.

Comme le numérateur converge en loi vers une normale centrée, et le dénominateur converge en probabilité vers une constante strictement positive, on peut appliquer le **théorème de Slutsky**, ce qui donne :

$$T \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Interprétation : La statistique du test de Welch converge donc asymptotiquement vers une loi normale standard, ce qui permet d'utiliser les quantiles de la loi normale pour les tests lorsque n_1 et n_2 sont grands.

4.3 Intervalles de confiance

Variances supposées égales

Lorsque les deux échantillons $X^{(1)} = (X_1^{(1)}, \ldots, X_{n_1}^{(1)})$ et $X^{(2)} = (X_1^{(2)}, \ldots, X_{n_2}^{(2)})$ sont indépendants et que l'on suppose que leurs variances sont égales, on utilise la variance combinée définie par :

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

où S_1^2 et S_2^2 sont les variances empiriques des deux échantillons.

L'intervalle de confiance exact au niveau de confiance $1-\alpha$ pour la différence de moyennes $\mu_1 - \mu_2$ est alors donné par :

$$\left[(\bar{X}^{(1)} - \bar{X}^{(2)}) \pm t_{1-\frac{\alpha}{2}, n_1+n_2-2} \cdot \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right],$$

où $t_{1-\frac{\alpha}{2}, n_1+n_2-2}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Lorsque n_1 et n_2 sont grands, on peut approcher la loi de Student par la loi normale standard, ce qui donne l'intervalle asymptotique suivant :

$$\left[(\bar{X}^{(1)} - \bar{X}^{(2)}) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right],$$

où $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale $\mathcal{N}(0,1)$.

Variances inégales (test de Welch)

Dans le cas où les variances ne sont pas supposées égales, on utilise l'intervalle de confiance suivant, exact au niveau $1-\alpha$:

$$\left[(\bar{X}^{(1)} - \bar{X}^{(2)}) \pm t_{1-\frac{\alpha}{2},\nu} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right],$$

avec les degrés de liberté ν approximés par la formule de Welch–Satterthwaite

L'intervalle asymptotique, lorsque les tailles d'échantillons sont grandes, est donné par :

$$\left[(\bar{X}^{(1)} - \bar{X}^{(2)}) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right].$$

4.4 Implémentation python

Dans cet exemple, nous comparons les moyennes d'âge entre les hommes et les femmes présents dans le jeu de données *Titanic*. Pour cela, nous utilisons un **test de Welch**. Ce test permet de déterminer si la différence observée entre les moyennes d'âge des deux groupes est statistiquement significative, sans supposer que les variances des deux populations sont égales. Cette approche est particulièrement adaptée dans le cas où les tailles d'échantillons ou les dispersions diffèrent entre les groupes.

```python
import pandas as pd
from scipy import stats

# Charger le dataset Titanic
titanic = pd.read_csv("titanic.csv")

# Supprimer les lignes avec des valeurs manquantes pour 'age' et 'sex'
titanic = titanic.dropna(subset=["Age", "Sex"])

# Extraire les deux échantillons
age_hommes = titanic[titanic["Sex"] == "male"]["Age"]
age_femmes = titanic[titanic["Sex"] == "female"]["Age"]


# Affichage des tailles
print(f"Moyenne d'age des hommes : {age_hommes.mean()}")
print(f"Moyenne d'age des femmes : {age_femmes.mean()}")

# Test t de Student à deux échantillons avec variances inégales (Welch)
t_stat, p_value = stats.ttest_ind(age_hommes, age_femmes, equal_var=False)

# Affichage des résultats
print(f"Statistique de test t : {t_stat:.4f}")
print(f"P-valeur : {p_value:.4f}")

# Décision au seuil de 5%
alpha = 0.05
if p_value < alpha:
    print("On rejette H0 : les moyennes d'âge sont significativement différentes.")
else:
    print("On ne rejette pas H0 : les moyennes d'âge peuvent être considérées comme égales.")
```

Figure 9 – Code test de Welch

```
Moyenne d'age des hommes : 30.72664459161148
Moyenne d'age des femmes : 27.915708812260537
Statistique de test t : 2.5259
P-valeur : 0.0118
On rejette H0 : les moyennes d'âge sont significativement différentes.
```

Figure 10 – Output du test

5 Test de Fisher pour la comparaison de deux variances

5.1 Objectif

On souhaite tester l'égalité de deux variances à partir de deux échantillons indépendants supposés suivre une loi normale. Soient :

$$X_1, \ldots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma_1^2), \quad Y_1, \ldots, Y_{n_2} \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

avec les deux échantillons indépendants.
On souhaite tester l'hypothèse :

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{contre} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

5.2 Définition de la loi de Fisher

Soient $U \sim \chi^2(d_1)$, $V \sim \chi^2(d_2)$, deux variables indépendantes suivant une loi du chi-deux avec respectivement d_1 et d_2 degrés de liberté. On définit la variable :

$$F = \frac{U/d_1}{V/d_2}$$

Alors F suit une loi de Fisher à d_1 et d_2 degrés de liberté, notée :

$$F \sim \mathcal{F}(d_1, d_2)$$

5.3 Statistique de test

On considère les estimateurs empiriques des variances :

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2, \quad S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_j - \bar{Y})^2$$

La statistique de test est définie par :

$$F = \frac{S_1^2}{S_2^2}$$

5.4 Comportement de la statistique sous H_0

Sous l'hypothèse H_0, on a $\sigma_1^2 = \sigma_2^2 = \sigma^2$. De plus, sous l'hypothèse de normalité :

$$(n_1 - 1)\frac{S_1^2}{\sigma^2} \sim \chi^2(n_1 - 1), \quad (n_2 - 1)\frac{S_2^2}{\sigma^2} \sim \chi^2(n_2 - 1)$$

et les deux variables sont indépendantes car les échantillons le sont.
On définit :

$$U = \frac{(n_1 - 1)S_1^2}{\sigma^2}, \quad V = \frac{(n_2 - 1)S_2^2}{\sigma^2}$$

Ainsi, la statistique peut s'écrire :

$$F = \frac{S_1^2}{S_2^2} = \frac{U/(n_1 - 1)}{V/(n_2 - 1)}$$

Par définition de la loi de Fisher, cela implique :

$$F \sim \mathcal{F}(n_1 - 1, n_2 - 1)$$

5.5 Conclusion

Sous H_0, si les échantillons sont indépendants et suivent une loi normale, alors la statistique de test :

$$F = \frac{S_1^2}{S_2^2}$$

suit une loi de Fisher $\mathcal{F}(n_1 - 1, n_2 - 1)$. Cette propriété permet d'utiliser la loi de Fisher pour effectuer un test d'égalité des variances.

5.6 Implémentation Python :

Ce test de Fisher permet de comparer les variances des âges entre les groupes hommes et femmes du dataset *Titanic*.

```python
import seaborn as sns
from scipy import stats
import pandas as pd

# Charger le dataset Titanic depuis seaborn
df = sns.load_dataset("titanic")

# Supprimer les lignes avec valeurs manquantes dans 'age' ou 'sex'
df = df.dropna(subset=["age", "sex"])

# Séparer les groupes
male_ages = df[df["sex"] == "male"]["age"]
female_ages = df[df["sex"] == "female"]["age"]

# Calcul des variances
var_male = male_ages.var(ddof=1)
var_female = female_ages.var(ddof=1)

# Statistique F
F_statistic = var_male / var_female

# Degrés de liberté
dof1 = len(male_ages) - 1
dof2 = len(female_ages) - 1

# p-value bilatérale
p_value = 2 * min(
    stats.f.cdf(F_statistic, dof1, dof2),
    1 - stats.f.cdf(F_statistic, dof1, dof2)
)

# Affichage des résultats
print(f"Statistique F : {F_statistic:.4f}")
print(f"p-value : {p_value:.4f}")

alpha = 0.05
if p_value < alpha:
    print("On rejette H0 : les variances d'âge sont significativement \
différentes entre hommes et femmes.")
else:
    print("On ne rejette pas H0 : les variances d'âge peuvent\
être considérées comme égales entre hommes et femmes.")
```

Figure 11 – Test de Fischer

```
Statistique F : 1.0821
p-value : 0.4814
On ne rejette pas H0 : les variances d'âge peuventêtre considérées comme égales entre hommes et femmes.
```

Figure 12 – Output

6 Test du χ^2 d'indépendance (variables qualitatives)

6.1 Objectif du test

Ce test permet de déterminer s'il existe une dépendance statistique entre deux variables qualitatives, en comparant les fréquences observées dans un tableau de contingence à celles attendues sous l'hypothèse d'indépendance. Il est très utilisé en analyse de données catégorielles, notamment en sciences sociales, santé publique ou finance.

6.2 Hypothèses

— H_0 : Les deux variables sont statistiquement indépendantes.
— H_1 : Les deux variables sont statistiquement dépendantes.

6.3 Conditions d'application

— Les données proviennent d'un échantillon aléatoire.
— Les observations sont indépendantes les unes des autres.
— Les effectifs théoriques attendus sont suffisamment grands : au moins 80% des cases doivent avoir un effectif ≥ 5, aucune case ne doit avoir un effectif < 1 (règle empirique classique).

6.4 Statistique de test

On considère un tableau de contingence à r lignes et c colonnes, où :
— r est le nombre de modalités de la première variable.
— c est le nombre de modalités de la seconde variable.

Soit O_{ij} l'effectif observé dans la cellule (i, j), et E_{ij} l'effectif attendu sous H_0, calculé par :

$$E_{ij} = \frac{(\text{total ligne } i) \times (\text{total colonne } j)}{\text{total général}}$$

La statistique de test est :

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

6.5 Loi de la statistique sous H_0

Sous l'hypothèse d'indépendance (H_0), la statistique χ^2 suit approximativement une loi du χ^2 à $(r-1)(c-1)$ degrés de liberté :

$$\chi^2 \sim \chi^2_{(r-1)(c-1)}$$

Ces degrés de liberté correspondent au nombre de valeurs indépendantes que peut prendre le tableau de contingence une fois les totaux marginaux fixés.

6.6 Décision

On fixe un niveau de signification α (généralement 5%), puis on détermine la valeur critique $\chi^2_{\alpha,(r-1)(c-1)}$ à partir des tables de la loi du χ^2.

— Si $\chi^2_{\text{observé}} > \chi^2_{\alpha,(r-1)(c-1)}$, on **rejette** H_0 : il existe une dépendance significative entre les deux variables.

— Sinon, on **ne rejette pas** H_0 : aucune preuve statistique d'association.

7 Exemple du test d'indépence du χ^2 pour 2 variables

On souhaite tester si deux variables qualitatives, ici le **sexe** (homme/femme) et la **boisson préférée au petit déjeuner** (café, matcha, jus d'orange), sont statistiquement indépendantes. Il s'agit donc d'un test d'indépendance basé sur une table de contingence.

7.1 Hypothèses

— H_0 : les deux variables sont indépendantes.

— H_1 : les deux variables sont dépendantes.

7.2 Table de contingence (effectifs observés)

Sexe	**Café**	**Matcha**	**Jus d'orange**	**Total**
Hommes	40	30	10	80
Femmes	80	20	20	120
Total	120	50	30	200

Table 3 – Table de contingence des effectifs observés

7.3 Effectifs attendus sous H_0

Sous l'hypothèse d'indépendance entre les deux variables (ici le sexe et la préférence de boisson), les effectifs théoriques attendus dans chaque cellule du tableau de contingence sont donnés par la formule :

$$E_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$$

où :

— E_{ij} est l'effectif attendu dans la cellule de la $i^{\text{ème}}$ ligne et de la $j^{\text{ème}}$ colonne,

— $n_{i\cdot}$ est le total de la $i^{\text{ème}}$ ligne (somme des observations pour le niveau i de la première variable),

— $n_{\cdot j}$ est le total de la $j^{\text{ème}}$ colonne (somme des observations pour le niveau j de la deuxième variable),

— n est le total général (somme de tous les effectifs observés).

La table des effectifs attendus est alors la suivante :

Sexe	**Café**	**Matcha**	**Jus d'orange**	**Total**
Hommes	48	20	12	80
Femmes	72	30	18	120
Total	120	50	30	200

Table 2 : Table des effectifs attendus sous l'hypothèse d'indépendance.

7.4 Statistique de test

La statistique du test est :

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(40-48)^2}{48} + \frac{(30-20)^2}{20} + \frac{(10-12)^2}{12} + \frac{(80-72)^2}{72} + \frac{(20-30)^2}{30} + \frac{(20-18)^2}{18}$$

$$\chi^2 = \frac{64}{48} + \frac{100}{20} + \frac{4}{12} + \frac{64}{72} + \frac{100}{30} + \frac{4}{18} \approx 1.33 + 5.00 + 0.33 + 0.89 + 3.33 + 0.22 = \boxed{11.1}$$

7.5 Degrés de liberté

Le nombre de degrés de liberté est :

$$\text{ddl} = (r-1)(c-1) = (2-1)(3-1) = \boxed{2}$$

où $r = 2$ (lignes : hommes/femmes) et $c = 3$ (colonnes : 3 types de boisson).

7.6 p-valeur et décision

Sous H_0, la statistique suit une loi du χ^2 à 2 degrés de liberté. La p-valeur est :

$$p = \mathbb{P}(\chi^2 \geq 11.1) = 1 - F_{\chi^2}(11.1; \text{ddl} = 2) \approx \boxed{0.0039}$$

Avec un seuil de signification $\alpha = 0{,}05$, on a :

$$p = 0.0039 < \alpha = 0.05 \Rightarrow \textbf{on rejette } H_0$$

7.7 Conclusion

Il existe une dépendance statistiquement significative entre le sexe et la boisson préférée au petit déjeuner. En d'autres termes, le choix de boisson semble influencé par le genre.

7.8 Implémentation en Python

```python
import numpy as np

observed = np.array([[40,30,10], [80,20,20]])

row_totals = np.sum(observed, axis=1)
col_totals = np.sum(observed, axis=0)
total = np.sum(observed)

#Compute expected values
expected = np.outer(row_totals, col_totals) / total

print(f'Table of expected values : \n{expected} ')

#chi square statistic

chi_square_statistic = np.sum((observed - expected)**2 / expected)

print(f'Value of the chi-square statistic : {chi_square_statistic}')

#degrees of freedom

degrees_of_freedom = (observed.shape[0] - 1) * (observed.shape[1] - 1)

p_value = 1 - stats.chi2.cdf(chi_square_statistic, degrees_of_freedom)

print( f'p-value : {p_value}')

print("Decision: ")

if p_value < 0.05:
    print('Dependent (reject H0)')
else:
    print('Independent (H0 holds true)')
```

Figure 13 – Implémentation manuelle du test de chi-2 d'indépendence

```
Table of expected values :
[[48. 20. 12.]
 [72. 30. 18.]]
Value of the chi-square statistic : 11.11111111111111
p-value : 0.003865920139472845
Decision:
Dependent (reject H0)
```

Figure 14 – Output

8 Analyse de variance (ANOVA) à 1 facteur

8.1 Objectif du test

L'analyse de variance (ANOVA) a pour objectif de tester l'**égalité des moyennes de plusieurs groupes**. Autrement dit, elle vérifie s'il existe une **différence significative** entre les groupes en comparant la variance **inter-groupes** à la variance **intra-groupes**.

8.2 Hypothèses

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1 : \exists\ i \neq j \text{ tel que } \mu_i \neq \mu_j \end{cases}$$

8.3 Conditions d'application

— Les observations sont i.i.d. et proviennent de *populations normales*.
— Les variances sont homogènes (équivalentes) — homoscédasticité.
— Les groupes sont indépendants les uns des autres.

8.4 Statistique de test

Soit :

$$SCE = \sum_i n_i (\bar{X}_i - \bar{X})^2,$$

la *somme des carrés des écarts entre groupes*, et :

$$SCI = \sum_i \sum_j (X_{ij} - \bar{X}_i)^2,$$

la *somme des carrés des écarts intragroupe*. Ici :

$$\bar{X}_i = \frac{1}{n_i} \sum_j X_{ij},$$

et :

$$\bar{X} = \frac{1}{N} \sum_i \sum_j X_{ij},$$

avec $N = \sum_i n_i$.

8.5 Statistic F de Fisher

$$F = \frac{SCE/(k-1)}{SCI/(N-k)},$$

où :

$$SCE/(k-1) \quad \text{est la variance inter-groupes,}$$
$$SCI/(N-k) \quad \text{est la variance intra-groupes,}$$

et k est le nombre de groupes.

8.6 Justification de la loi de la statistique de test dans l'ANOVA à un facteur

On considère k échantillons indépendants :

$$X_{i1}, X_{i2}, \ldots, X_{in_i} \sim \mathcal{N}(\mu_i, \sigma^2), \quad \text{avec } i = 1, \ldots, k$$

où chaque échantillon est composé d'observations i.i.d. issues d'une loi normale de variance commune σ^2.

On note :
- n_i la taille du i-ème groupe ;
- $N = \sum_{i=1}^{k} n_i$ la taille totale ;
- $\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$ la moyenne du groupe i ;
- $\bar{X} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}$ la moyenne globale.

Hypothèses du test

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu, \qquad H_1 : \exists i \neq j \text{ tel que } \mu_i \neq \mu_j$$

Décomposition de la variance On définit les trois sommes des carrés suivantes :
- **Somme des carrés totale (SCT)** :

$$\text{SCT} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$$

- **Somme des carrés entre groupes (SCE)** :

$$\text{SCE} = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X})^2$$

- **Somme des carrés intra-groupes (SCI)** :

$$\text{SCI} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

La relation fondamentale est donnée par l'identité de Huygens :

$$\text{SCT} = \text{SCE} + \text{SCI}$$

Statistique de test On définit la statistique de test :

$$F = \frac{\text{SCE}/(k-1)}{\text{SCI}/(N-k)}$$

Loi de la statistique sous H_0 Sous l'hypothèse nulle H_0, toutes les observations X_{ij} suivent une loi normale centrée sur une même moyenne μ, donc $X_{ij} \sim \mathcal{N}(\mu, \sigma^2)$.

Il est alors possible de démontrer que :

1. La quantité $\dfrac{\text{SCI}}{\sigma^2}$ suit une loi du χ^2 à $N-k$ degrés de liberté :

$$\frac{\text{SCI}}{\sigma^2} \sim \chi^2(N-k)$$

   Cela résulte du fait que SCI est une somme de carrés de $N-k$ variables normales centrées indépendantes. L'utilisation des moyennes de groupes (au nombre de k) impose k contraintes linéaires sur les N observations, donc il reste $N-k$ degrés de liberté (démonstration en annexe).

2. De même, la quantité $\dfrac{\text{SCE}}{\sigma^2}$ suit une loi du χ^2 à $k-1$ degrés de liberté :

$$\frac{\text{SCE}}{\sigma^2} \sim \chi^2(k-1)$$

   car elle résulte de la somme des carrés des différences entre les k moyennes de groupes $\bar{X}_i$ et la moyenne globale $\bar{X}$, avec une seule contrainte (la somme pondérée des écarts est nulle), ce qui donne $k-1$ degrés de liberté.

3. Ces deux composantes SCE et SCI sont issues de projections orthogonales dans un espace euclidien, donc elles sont indépendantes :

$$\text{SCE} \perp \text{SCI}$$

Par définition de la loi de Fisher :

$$\text{Si } U \sim \chi^2(d_1), \quad V \sim \chi^2(d_2), \quad \text{avec } U \perp V, \quad \text{alors } \frac{U/d_1}{V/d_2} \sim \mathcal{F}(d_1, d_2)$$

On conclut donc que, sous H_0 :

$$F = \frac{\text{SCE}/(k-1)}{\text{SCI}/(N-k)} \sim \mathcal{F}(k-1, N-k)$$

□

8.7 Décision

Soit $F_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi de Fisher. Si :

$$F > F_{1-\alpha},$$

alors on **rejette** H_0. Sinon, on **conserve** H_0.

Exemple d'application de l'ANOVA

Afin d'illustrer l'utilisation du test d'Analyse de la Variance (ANOVA) à un facteur, nous appliquons cette méthode sur le jeu de données *Titanic*. L'objectif est ici de tester si l'âge moyen des passagers varie significativement selon leur classe sociale à bord (`Pclass`), c'est-à-dire de comparer les moyennes d'âge entre les classes 1, 2 et 3.

Le facteur étudié est donc la variable catégorielle `Pclass` (classe du billet), et la variable dépendante est l'âge des passagers. Nous nous intéressons à savoir si les différences d'âge entre les groupes sont statistiquement significatives, ou si elles peuvent être attribuées au hasard. Pour cela, nous réalisons une ANOVA à un facteur en Python, et nous commentons les résultats obtenus à l'aide de la statistique F et de la p-valeur associée.

```
[ ]  import seaborn as sns
     import pandas as pd
     from scipy import stats

     # Charger les données Titanic depuis seaborn
     df = sns.load_dataset("titanic")

     # Nettoyage : enlever les valeurs manquantes
     df_clean = df[["age", "pclass"]].dropna()

     # Groupes par classe
     group1 = df_clean[df_clean["pclass"] == 1]["age"]
     group2 = df_clean[df_clean["pclass"] == 2]["age"]
     group3 = df_clean[df_clean["pclass"] == 3]["age"]

     # Test ANOVA
     f_statistic, p_value = stats.f_oneway(group1, group2, group3)

     print(f"Statistique F : {f_statistic:.4f}")
     print(f"p-value : {p_value:.4f}")

     # Interprétation du test à un seuil de 5%
     alpha = 0.05
     if p_value < alpha:
         print("On rejette H0 : les moyennes d'âge diffèrent significativement selon la classe.")
     else:
         print("On ne rejette pas H0 : pas de différence significative entre les moyennes d'âge.")
```

Figure 15 – Implémentation de l'ANOVA sur l'age des passagers en fonction de leur classe

```
Statistique F : 57.4435
p-value : 0.0000
On rejette H0 : les moyennes d'âge diffèrent significativement selon la classe.
```

Figure 16 – Résultat du test

Conclusion de l'analyse La statistique de test obtenue est $F = 57{,}4435$, avec une p-valeur associée extrêmement faible (inférieure à 10^{-4}). Ce résultat est largement significatif au seuil de 5%. Ainsi, nous rejetons l'hypothèse nulle d'égalité des moyennes : il existe une différence significative des âges moyens entre les différentes classes de passagers du Titanic. Ce résultat est cohérent avec l'idée que les passagers des classes sociales supérieures étaient en moyenne plus âgés que ceux des classes inférieures comme le montre ces moyennes.

```python
print("Moyenne d'âge de la classe 1 :", round(group1.mean(), 2))
print("Moyenne d'âge de la classe 2 :", round(group2.mean(), 2))
print("Moyenne d'âge de la classe 3 :", round(group3.mean(), 2))
```

```
Moyenne d'âge de la classe 1 : 38.23
Moyenne d'âge de la classe 2 : 29.88
Moyenne d'âge de la classe 3 : 25.14
```

Figure 17 – Résultat du test

9 Minimax, erreurs de type I/II et taux de convergence

Cadre et motivation

En statistique non paramétrique, l'objet que l'on cherche à estimer ou tester n'est pas un paramètre unique (comme une moyenne ou une variance), mais peut être une fonction entière, par exemple une densité de probabilité. On ne suppose donc pas que les données suivent un modèle simple de dimension finie.

On considère typiquement un test du type :

$$H_0 : P \in \mathcal{P}_0 \quad \text{vs} \quad H_1 : P \in \mathcal{P}_1,$$

où $\mathcal{P}_0$ et $\mathcal{P}_1$ sont des classes larges de lois possibles sur l'espace des observations.

Dans ce contexte, chercher à maximiser la puissance contre une seule alternative ponctuelle n'a pas beaucoup de sens, car il y a une infinité de façons dont la vraie loi peut s'écarter de $\mathcal{P}_0$. C'est pourquoi on adopte une approche *minimax* : on construit des tests qui contrôlent uniformément le risque maximal sur toutes les lois de $\mathcal{P}_0$ et $\mathcal{P}_1$, en garantissant de bonnes performances même dans le pire cas.

Cette démarche est spécifique à la statistique non paramétrique. Dans un cadre paramétrique classique, l'alternative est souvent ponctuelle (un paramètre précis), et on peut directement optimiser la puissance contre cette alternative. En non paramétrique, le nombre infini de possibles alternatives rend cette stratégie impossible.

Erreurs de type I et II uniformisées et risque maximal

Pour un test ϕ (une fonction qui prend 1 si on rejette H_0 et 0 sinon), on définit :

$$\alpha(\phi) = \sup_{P \in \mathcal{P}_0} P(\phi = 1) \quad \text{(erreur de type I maximale)},$$

$$\beta(\phi) = \sup_{Q \in \mathcal{P}_1} Q(\phi = 0) \quad \text{(erreur de type II maximale)}.$$

On définit ensuite un *risque global* comme une combinaison de ces deux erreurs, par exemple :

$$R(\phi) = \alpha(\phi) + \beta(\phi).$$

L'approche *minimax* consiste à construire des tests qui contrôlent ce risque de manière uniforme sur toutes les lois de $\mathcal{P}_0$ et $\mathcal{P}_1$, et à identifier le meilleur taux possible lorsque la taille de l'échantillon n augmente.

Définition : taux de séparation minimax

Soient (r_n) et (ρ_n) deux suites de réels positifs ou nuls, et soit $H_0 \subset \mathcal{F}$ une hypothèse statistique. Soit d une métrique sur $\mathcal{F}$. La suite $(\rho_n : n \in \mathbb{N})$ est appelée *taux de séparation minimax* pour le test

$$f \in H_0 \quad \text{contre} \quad f \in H_1 = H_1(d, r_n) = \left\{ f \in \mathcal{F} : \inf_{g \in H_0} d(f, g) \geq r_n \right\}$$

si les deux conditions suivantes sont vérifiées :

(i) Pour tout $\alpha > 0$, il existe un test φ_n tel que, pour n assez grand,

$$\sup_{f \in H_0} \mathbb{E}_f[\varphi_n] \; + \sup_{f \in H_1(d,\rho_n)} \mathbb{E}_f[1 - \varphi_n] \; \leq \; \alpha.$$

(ii) Pour toute suite $r_n = o(\rho_n)$, on a

$$\liminf_{n \to \infty} \; \inf_{\varphi_n} \left\{ \sup_{f \in H_0} \mathbb{E}_f[\varphi_n] \; + \sup_{f \in H_1(d,r_n)} \mathbb{E}_f[1 - \varphi_n] \right\} \; > \; 0,$$

où l'infimum est pris sur toutes les fonctions mesurables $\varphi_n : \mathcal{Y}_n \to \{0, 1\}$.
Dans le cas du test de Kolmogorov–Smirnov, la métrique considérée est

$$d(f, g) = \|F_f - F_g\|_\infty,$$

où F_f et F_g désignent les fonctions de répartition associées aux densités f et g.

Application : taux de séparation minimax du test de Kolmogorov–Smirnov pour la loi uniforme sur $[0,1]$

Cadre Dans ce qui suit, nous établissons le meilleur taux de séparation minimax pour le test de Kolmogorov–Smirnov lorsque la loi nulle est la loi uniforme sur $[0,1]$. Soit $X_1, \dots, X_n$ un échantillon i.i.d. de loi de répartition F. On considère le test des hypothèses

$$H_0 : F(t) = t \quad \forall t \in [0,1] \qquad \text{vs.} \qquad H_1 : F \in \Big\{ \sup_{t\in[0,1]} |F(t) - t| \geq r_n \Big\}.$$

Nous montrerons, au sens de la définition précédente, que le taux de séparation minimax est

$$\rho_n = \frac{C}{\sqrt{n}},$$

où la constante $C > 0$ dépend uniquement du niveau asymptotique α du test.

Borne supérieure :

Statistique et test. Soit F_n la fonction de répartition empirique construite à partir de $X_1, \dots, X_n$. Considérons la statistique de Kolmogorov–Smirnov standard

$$T_n \;=\; \sqrt{n} \sup_{t\in[0,1]} |F_n(t) - F_0(t)| \;=\; \sqrt{n}\, \|F_n - F_0\|_\infty,$$

où $F_0(t) = t$ est la loi uniforme sur $[0,1]$. Pour un seuil $z > 0$ (que nous choisirons ci-dessous), posons le test

$$\varphi_n \;=\; \mathbf{1}\{T_n > z\} \;=\; \mathbf{1}\Big\{ \|F_n - F_0\|_\infty > \frac{z}{\sqrt{n}} \Big\}.$$

But. Nous voulons montrer qu'il existe une constante $C > 0$ (fonction de α seulement) et un choix de z tels que, pour $\rho_n = C/\sqrt{n}$,

$$\sup_{F\in H_0} \mathbb{E}_F[\varphi_n] \;+\; \sup_{F\in H_1(\|\cdot\|_\infty, \rho_n)} \mathbb{E}_F[1 - \varphi_n] \;\leq\; \alpha \quad \text{pour tout } n \text{ assez grand.}$$

Contrôle de l'erreur de type I (sous H_0). Sous H_0 (ici F_0 fixe), par le théorème de Donsker,

$$T_n = \sqrt{n}\|F_n - F_0\|_\infty \;\xrightarrow[n\to\infty]{}\; \|G\|_\infty,$$

où G est un pont brownien sur $[0,1]$ (on l'admettra ici). Soit $z_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi de $\|G\|_\infty$. En choisissant $z = z_{1-\alpha}$, on obtient, par convergence en loi,

$$\limsup_{n\to\infty} \sup_{F\in H_0} P_F(T_n > z) = P(\|G\|_\infty > z) = \alpha.$$

Donc, pour n assez grand,

$$\sup_{F\in H_0} \mathbb{E}_F[\varphi_n] = \sup_{F\in H_0} P_F(T_n > z) \leq \alpha.$$

Contrôle de l'erreur de type II. L'erreur de type II est la probabilité de ne pas rejeter H_0 sous F :

$$\mathbb{E}_F[1-\varphi_n] = \mathbb{P}_F(T_n \le z_\alpha) = \mathbb{P}_F\Big(\sqrt{n}\,\|F_n - F_0\|_\infty \le z_\alpha\Big).$$

En utilisant l'inégalité triangulaire :

$$\|F_n - F_0\|_\infty \ge \|F - F_0\|_\infty - \|F_n - F\|_\infty,$$

on obtient

$$\sqrt{n}\,\|F_n - F_0\|_\infty \ge C - \sqrt{n}\,\|F_n - F\|_\infty.$$

car on suppose $F \in H_1(d, \rho_n)$, comme dans l'inégalité (i) , avec $\rho_n = C/\sqrt{n}$, donc $\|F - F_0\|_\infty < C/\sqrt{n}$. .

Ainsi, on obtient la majoration suivante de l'erreur de type 2 :

$$\mathbb{E}_F[1-\varphi_n] = \mathbb{P}_F\Big(\sqrt{n}\,\|F_n - F_0\|_\infty \le z_\alpha\Big) \le \mathbb{P}_F\Big(\sqrt{n}\,\|F_n - F\|_\infty \ge C - z_\alpha\Big).$$

Par le théorème de Donsker ou l'inégalité de Dvoretzky–Kiefer–Wolfowitz, $\sqrt{n}(F_n - F)$ converge en loi vers un pont brownien G_F, ce qui permet de choisir C suffisamment grand pour que

$$\mathbb{E}_F[1-\varphi_n] \le \beta,$$

tant que $C - z_\alpha \ge z_\beta$, avec β arbitrairement petit. Ainsi, pour des alternatives suffisamment séparées, l'erreur de type II est contrôlée.

Conclusion. En réunissant les deux contrôles, on obtient, pour n assez grand,

$$\sup_{F \in H_0} \mathbb{E}_F[\varphi_n] + \sup_{F \in H_1(\|\cdot\|_\infty, \rho_n)} \mathbb{E}_F[1-\varphi_n] \le \alpha + \beta,$$

où $\varphi_n = \mathbf{1}\{\sqrt{n}\,\|F_n - F_0\|_\infty > z_\alpha\}$. En choisissant β petit, cette somme peut être rendue inférieure à un niveau arbitraire $\alpha > 0$.
Ainsi, la condition (i) de la définition est vérifiée avec $\rho_n = C/\sqrt{n}$.

Remarque : La preuve présentée ici est générale dans le cas du test de Kolmogorov Smirnov et ne tient pas compte du fait que nous considérons spécifiquement le test de KS pour la loi uniforme sur $[0, 1]$.

Borne inférieure :

Avant de commencer la preuve de l'inégalité (ii) de la Définition, rappelons une borne inférieure qu'on ne démontrera pas mais qui sera utile par la suite.

Sous certaines conditions (qu'on respecte dans le cas de notre test de KS sur la loi uniforme), pour tout $\eta > 0$, on a

$$\inf_{\varphi} \left\{ \mathbb{E}_{f_0}[\varphi] + \sup_{f \in \mathcal{M}} \mathbb{E}_f[1-\varphi] \right\} \geq (1-\eta) \left(1 - \frac{\sqrt{\mathbb{E}_{f_0}(Z-1)^2}}{\eta} \right) \tag{1}$$

où

$$Z = \frac{1}{M} \sum_{m=1}^{M} \frac{dP_{f_m}}{dP_{f_0}}$$

est la moyenne des rapports de vraisemblance.

Cadre. Soit $f_0 = 1$ la densité uniforme sur $[0,1]$ et soit ψ une fonction bornée, à support dans $[0,1]$, telle que

$$\|\psi\|_1 \leq \|\psi\|_2 = 1 \quad \text{et} \quad \int_0^1 \psi(x)\,dx = 0.$$

Construction de l'alternative. Pour $r_n = o(n^{-1/2})$, posons $\psi_n = r_n \psi$ et

$$f_1 = f_0 + \psi_n.$$

Pour n assez grand (selon $\|\psi\|_\infty$), f_1 reste positive et définit une densité. Sa fonction de répartition F_1 satisfait

$$\|F_0 - F_1\|_\infty = r_n \sup_{t \in [0,1]} \left| \int_0^t \psi(x)\,dx \right| = C r_n,$$

et donc $F_1 \in H_1(r_n)$ dès que n est assez grand.

Application de la minoration de (ii). On applique l'inégalité énoncée en préambule de la preuve avec $M = 1$ (M est la cardinal de la famille d'alternatives, donc ici 1 car on test une seule alternative) : il suffit de contrôler la distance χ^2 entre P_{f_1} et P_{f_0}. Pour

$$Z = \frac{dP_{f_1}}{dP_{f_0}},$$

il faut donc évaluer $E_{f_0}[(Z-1)^2]$.

Calcul de $E_{f_0}[(Z-1)^2]$. Sous P_{f_0} (densité $f_0 = 1$),

$$Z(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_1(x_i) = \prod_{i=1}^{n} \big(1 + r_n \psi(x_i)\big).$$

On a alors

$$E_{f_0}[(Z-1)^2] = \int_{[0,1]^n} \Big(\prod_{i=1}^n f_1(x_i) - 1 \Big)^2 dx_1 \dots dx_n$$
$$= \int_{[0,1]^n} \prod_{i=1}^n (1 + r_n \psi(x_i))^2 \, dx_1 \dots dx_n - 1,$$

où la deuxième égalité vient du fait que

$$\int_{[0,1]^n} \prod_{i=1}^n f_1(x_i) \, dx = 1$$

puisque f_1 est une densité.

Par indépendance, l'intégrale se factorise :

$$\int_{[0,1]^n} \prod_{i=1}^n (1 + r_n \psi(x_i))^2 \, dx_1 \dots dx_n = \left(\int_0^1 (1 + r_n \psi(x))^2 dx \right)^n.$$

Or

$$\int_0^1 (1 + r_n \psi(x))^2 dx = \int_0^1 \big(1 + 2 r_n \psi(x) + r_n^2 \psi(x)^2\big) \, dx = 1 + r_n^2,$$

car $\int_0^1 \psi(x) \, dx = 0$ et $\|\psi\|_2 = 1$. Ainsi

$$E_{f_0}[(Z-1)^2] = (1 + r_n^2)^n - 1.$$

En utilisant l'inégalité $1 + x \leq e^x$, on obtient

$$(1 + r_n^2)^n - 1 \;\leq\; e^{n r_n^2} - 1.$$

Comme $r_n = o(n^{-1/2})$, on a $n r_n^2 \to 0$, donc

$$E_{f_0}[(Z-1)^2] \;\longrightarrow\; 0.$$

Conclusion L'inégalité (1) montre que si $E_{f_0}[(Z-1)^2] \to 0$, alors pour $\eta \in]0,1[$ (ça n'a pas d'utilité de le prendre plus grand que 1 car la partie gauche est forcément positive), on a bien une minoration du risque (ii) strictement positive.

Autrement dit, si $r_n = o(n^{-1/2})$, alors pour **tout test** φ_n (donc pas uniquement KS), la somme des erreurs

$$\sup_{f \in H_0} \mathbb{E}_f[\varphi_n] + \sup_{f \in H_1(d, r_n)} \mathbb{E}_f[1 - \varphi_n]$$

reste **strictement positive** pour n grand, c'est-à-dire que le risque ne peut pas tendre vers 0 uniformément sur les alternatives.

Conclusion sur le taux minimax

En combinant les résultats des bornes supérieure et inférieure, on peut conclure que le taux de séparation minimax pour le test de Kolmogorov–Smirnov sur la loi uniforme $\mathcal{U}[0,1]$ est

$$\rho_n = \frac{C}{\sqrt{n}},$$

où $C > 0$ dépend uniquement du niveau asymptotique α du test.

— **Borne supérieure :** Nous avons construit un test basé sur la statistique

$$T_n = \sqrt{n}\|F_n - F_0\|_\infty,$$

et montré qu'il existe un seuil z tel que

$$\sup_{F\in H_0} \mathbb{E}_F[\phi_n] + \sup_{F\in H_1(\|\cdot\|_\infty,\rho_n)} \mathbb{E}_F[1-\phi_n] \leq \alpha, \quad \alpha \geq 0$$

ce qui valide la condition (i) de la définition .

— **Borne inférieure :** En utilisant l'inégalité (1) et une alternative construite $f_1 = f_0 + r_n\psi$ avec $r_n = o(\rho_n)$, on obtient

$$\inf_{\varphi_n} \Big\{ \sup_{f\in H_0} \mathbb{E}_f[\varphi_n] + \sup_{f\in H_1(d,r_n)} \mathbb{E}_f[1-\varphi_n] \Big\} \geq \text{constante } > 0,$$

ce qui montre que la condition (ii) de la définition est respectée.

Ainsi, les deux conditions de la définition du taux minimax sont satisfaites et on en déduit que le taux minimax est de l'ordre

$$\rho_n \sim \frac{1}{\sqrt{n}}.$$

Toutes les définitions et preuves utilisées dans cette section, y compris la définition du taux de séparation minimax et les démonstrations des bornes supérieure et inférieure pour le test de Kolmogorov–Smirnov sur la loi uniforme, peuvent être consultées dans le livre de Giné et Nickl "Mathematical Foundations of Infinite-Dimensional Statistical Models" (Cambridge University Press), j'ai essayé de détailler et expliquer les calculs pour faciliter leur compréhension, et de mettre des mots sur les inégalités et les raisonnements, pour retranscrire la compréhension qu'il faut (je pense) en avoir.

10 Annexe

Dans cette annexe, nous regroupons quelques définitions et théorèmes fondamentaux qui interviennent dans les démonstrations et résultats présentés. Ces rappels ont pour but d'éclaircir certains passages sans alourdir le texte principal.

10.1 Principes de base des tests

Avant d'aborder des résultats plus techniques, rappelons brièvement quelques notions fondamentales utilisées en théorie des tests statistiques :

- **Hypothèses** :
  - L'hypothèse nulle H_0 correspond à la situation « de référence » que l'on cherche à tester.
  - L'hypothèse alternative H_1 regroupe les situations où H_0 est fausse.
- **Niveau (ou risque de première espèce)** : C'est la probabilité de rejeter H_0 alors qu'elle est vraie, généralement notée α. Un test est dit de *niveau* α si

$$\sup_{F \in H_0} \mathbb{P}_F(\text{rejeter } H_0) \leq \alpha.$$

- **Puissance du test** : C'est la probabilité de rejeter H_0 lorsque l'alternative H_1 est vraie. Elle est souvent notée $1-\beta$, où β est le risque de deuxième espèce.

- **Intervalle de confiance** : Un intervalle de confiance de niveau $1-\alpha$ est un ensemble aléatoire qui contient le vrai paramètre avec probabilité au moins $1-\alpha$. Il est étroitement lié aux tests : un paramètre est rejeté par un test bilatéral au niveau α s'il n'appartient pas à l'intervalle de confiance correspondant.

- **Fonction de test** : Un test peut être vu comme une fonction φ_n qui vaut 1 si l'on rejette H_0 et 0 sinon. Les propriétés du test s'expriment alors par des conditions sur $\mathbb{E}_F[\varphi_n]$ (probabilité de rejet sous F).

10.2 Lois de Student et du χ^2 : rappels et liens

Dans de nombreuses procédures de test, apparaissent la loi du χ^2 et la loi de Student, que nous rappelons brièvement :

- **Loi du χ^2** : Si $Z_1, \ldots, Z_k$ sont indépendantes et suivent une loi normale centrée réduite $\mathcal{N}(0,1)$, alors

$$\chi^2_k = \sum_{i=1}^{k} Z_i^2$$

  suit une loi du χ^2 à k degrés de liberté.

— **Loi de Student** : Si $Z \sim \mathcal{N}(0,1)$ et $V \sim \chi^2_k$ sont indépendants, alors

$$T = \frac{Z}{\sqrt{V/k}}$$

suit une loi de Student à k degrés de liberté, notée t_k.

10.3 Théorèmes fondamentaux : TCL et théorème de Slutsky

Pour justifier la convergence en loi de nombreuses statistiques utilisées dans les tests, nous rappelons deux résultats centraux :

— **Théorème Central Limite (TCL)** : Soient $X_1, X_2, \ldots, X_n$ des variables aléatoires i.i.d. de moyenne μ et de variance finie $\sigma^2 > 0$. Alors, lorsque $n \to \infty$,

$$\sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma} \quad \xrightarrow{d} \quad \mathcal{N}(0,1),$$

où $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ et $\xrightarrow{d}$ désigne la convergence en loi.

— **Théorème de Slutsky** : Soient (X_n) et (Y_n) deux suites de variables aléatoires telles que

$$X_n \xrightarrow{d} X \quad \text{et} \quad Y_n \xrightarrow{p} c,$$

avec c une constante réelle. Alors :

$$X_n + Y_n \xrightarrow{d} X + c, \qquad X_n Y_n \xrightarrow{d} cX, \qquad \frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c} \ (c \neq 0).$$

Ce théorème permet de remplacer une quantité aléatoire qui converge en probabilité vers une constante par cette constante dans les limites en loi.