Cartographie des Tests Statistiques

Théorie, Applications et Implémentations

Abdelli Farès

Mai 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`.

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 :

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

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.

```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

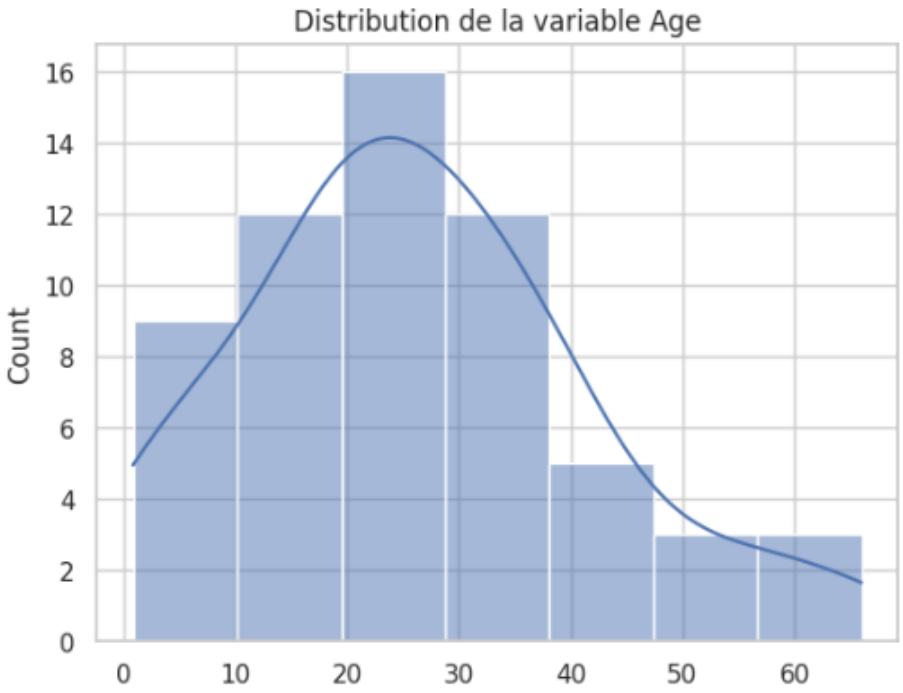

FIGURE 3 – KDE Shapiro

1.2 Test de Kolmogorov-Smirnov

Utilité du test

Le test de Kolmogorov-Smirnov (ou test KS) est un test non paramétrique utilisé pour vérifier si un échantillon provient d'une distribution spécifique, souvent une distribution continue comme la loi normale. Il permet de comparer la fonction de répartition empirique d'un échantillon à la fonction de répartition théorique d'une loi hypothétique. Ce test est particulièrement utile pour évaluer la conformité à une loi sans supposer de paramètres particuliers ou sans nécessiter que la variance soit connue.

—

Hypothèses du test

On considère un échantillon $\{X_1, X_2, \ldots, X_n\}$ de taille n, et une loi cumulative théorique $F_0(x)$. Le test a pour hypothèses :

$$\begin{cases} H_0 : \text{la distribution des } X_i \text{ est } F_0(x) \\ H_1 : \text{la distribution des } X_i \text{ n'est pas } F_0(x) \end{cases}$$

—

Formalisme mathématique et statistique de test

La statistique de test du test de Kolmogorov-Smirnov est basée sur la distance maximale entre la fonction de répartition empirique $F_n(x)$ de l'échantillon et la fonction de répartition théorique $F_0(x)$.

La fonction de répartition empirique est définie par :

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

où $\mathbf{1}_{X_i \leq x}$ est la fonction indicatrice valant 1 si $X_i \leq x$ et 0 sinon.

La statistique de test est alors :

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|$$

où sup est la borne supérieure (le maximum) sur tous les x.

Intuitivement, D_n mesure la plus grande différence absolue entre la courbe empirique et la courbe théorique.

—

Règle de décision

La distribution de la statistique D_n sous l'hypothèse nulle H_0 ne dépend pas de la distribution F_0 (si elle est continue) : cette propriété rend le test de Kolmogorov-Smirnov universellement applicable.

Pour un niveau de confiance α, on compare D_n à une valeur critique $c(\alpha)$ issue de la table de la loi de Kolmogorov (ou calculée via des approximations asymptotiques) :

$$\text{On rejette } H_0 \text{ si } D_n > c(\alpha)$$

Sinon, on ne rejette pas H_0.

—

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.
- Le traitement du signal et des séries temporelles, notamment pour tester la stationnarité ou la distribution des innovations.

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.

```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

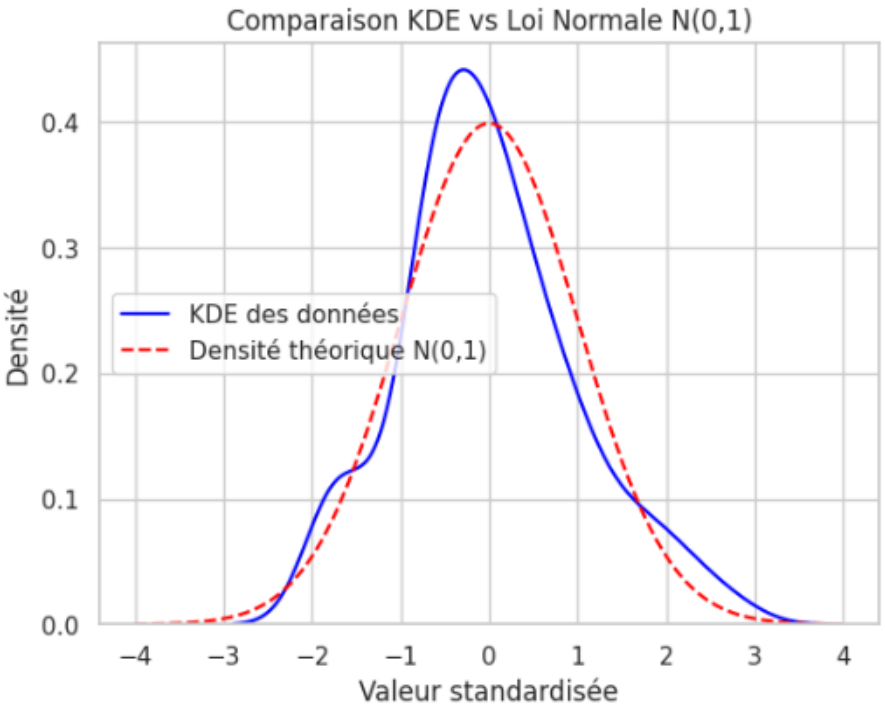

FIGURE 6 – KDE KS

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.

3 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.

```
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.
```

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

4.1 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.

4.2 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.

4.3 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.

4.4 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 :

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

4.5 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 $\bar{X}$. Ce résultat découle de la théorie classique des statistiques sur les lois normales.

4.6 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.

4.7 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 $\mathcal{N}(0, 1)$.

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

5.1 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 :

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

5.2 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 :

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

5.3 Indépendance de $\bar{X}$ et S^2

Un résultat fondamental 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.

5.4 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.

5.5 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.

5.6 Conclusion

La statistique T suit une loi de Student, non pas par hypothèse, mais parce qu'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.

—

5.7 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.

5.8 Implémentation python

```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

6 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.

Cas 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 :

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

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

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

Cas 2 : Variances inconnues et supposées différentes (test de Welch)

Dans ce cas, on ne fait plus l'hypothèse de variance commune.
Statistique de test :

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

Degrés de liberté effectifs : selon l'approximation de Welch–Satterthwaite :

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

Sous H_0, la statistique T suit approximativement une loi de Student à ν degrés de liberté.
Règle de décision :

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

Remarque : approximation par la loi normale

Lorsque $n_1, n_2 \geq 30$, on peut utiliser l'approximation suivante :

$$T \overset{approx}{\sim} \mathcal{N}(0, 1)$$

Et on utilise alors :

$$\text{On rejette } H_0 \text{ si } |T| > z_{1-\alpha/2}$$

Résumé

— **Égalité des variances** ⇒ test t classique avec $n_1 + n_2 - 2$ d.d.l.
— **Variances inégales** ⇒ test de Welch avec degrés de liberté effectifs.
— **Échantillons grands** ⇒ approximation normale possible.

6.1 Implémentation python

```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

7 Test de Fisher pour la comparaison de deux variances

Le test de Fisher permet de tester l'égalité de deux variances issues de deux échantillons indépendants supposés suivre une loi normale.

Hypothèses

Soient deux échantillons :

$$X_1, \ldots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$$
$$X_2, \ldots, X_{n_2} \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

On souhaite tester :

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

Statistique de test

Les variances empiriques sont données par :

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

La statistique de test est :

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

Sous l'hypothèse nulle H_0, F suit une loi de Fisher $\mathcal{F}(n_1 - 1, n_2 - 1)$.

Règle de décision

On fixe un seuil α (typiquement 0,05). On rejette H_0 si :

$$F < F_{\alpha/2; n_1-1, n_2-1} \quad \text{ou} \quad F > F_{1-\alpha/2; n_1-1, n_2-1}$$

Alternativement, on peut calculer la p-value du test. Si p-value $< \alpha$, alors on rejette H_0.

Remarque

Le test de Fisher suppose la normalité des deux échantillons. En cas de doute sur cette hypothèse, il est préférable d'utiliser un test robuste comme celui de Levene.

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

8.1 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.

8.2 Hypothèses

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

8.3 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).

8.4 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 :

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

8.5 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.

8.6 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.

8.7 Remarque pratique

Pour des tableaux de petite taille, notamment 2×2, ou lorsque les conditions d'application ne sont pas respectées, il est recommandé d'utiliser le *test exact de Fisher*, plus fiable dans ce cas. En Python, le test du χ^2 peut être implémenté avec `scipy.stats.chi2_contingency`.

9 Test du χ^2 d'indépendance : exemple

9.1 Exemple

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.

9.2 Hypothèses

— H_0 : les deux variables sont indépendantes.
— H_1 : les deux variables sont dépendantes.

9.3 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 1 – Table de contingence des effectifs observés

9.4 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 :

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

où :
— E_{ij} est l'effectif attendu dans la cellule de la $i^{\text{ème}}$ ligne et de la $j^{\text{ème}}$ colonne,
— $n_{i.}$ est le total de la $i^{\text{ème}}$ ligne (somme des observations pour le niveau i de la première variable),
— $n_{.j}$ 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.

9.5 Statistique de test

La statistique du test est :

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

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

9.6 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).

9.7 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 :

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

9.8 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.

9.9 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 11 – 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 12 – Output

10 Analyse de variance (ANOVA) à 1 facteur

10.1 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**.

10.2 Hypothèses

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

10.3 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.

10.4 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 $N = \sum_i n_i$.

10.5 Statistic F de Fisher

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

où :

$SCE/(k-1)$ est la variance inter-groupes,

$SCI/(N-k)$ est la variance intra-groupes,

et k est le nombre de groupes.

]Loi de la statistique sous H_0"
Sous l'hypothèse nulle H_0, le rapport F suit une *loi de Fisher-Snedecor* :

$$F \sim F(k-1, N-k).$$

10.6 Décision

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

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

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

11 Test de Dickey-Fuller (simple) sur les racines unitaires

11.1 Idée du test

Le test de Dickey-Fuller simple consiste à tester l'existence d'une racine unitaire dans une série chronologique. Autrement dit, on vérifie si le choc aléatoire a un *effet permanent* sur le système (série non-stationnaire) ou s'il s'estompe avec le temps (série stationnaire).

11.2 Modèle

Le modèle de base de Dickey-Fuller est :

$$Y_t = \rho \cdot Y_{t-1} + \epsilon_t,$$

où :

- Y_t est la valeur de la série au temps t.
- ρ est le facteur autorégressif.
- ϵ_t est un bruit blanc iid de moyenne nulle et de variance constante.

11.3 Transformations

En posant $\delta = \rho - 1$, le modèle peut s'écrire :

$$\Delta Y_t = \delta \cdot Y_{t-1} + \epsilon_t,$$

où :

$$\Delta Y_t = Y_t - Y_{t-1}.$$

11.4 Hypothèses

$$\begin{cases} H_0 : \delta = 0 & (\rho = 1 \quad \rightarrow \quad \text{Série non-stationnaire}) \\ H_1 : \delta < 0 & (\rho < 1 \quad \rightarrow \quad \text{Série stationnaire}) \end{cases}$$

11.5 Statistique de test

L'estimateur des moindres carrés de δ est :

$$\hat{\delta} = \frac{\sum_{t=2}^{T}(Y_{t-1} \cdot \Delta Y_t)}{\sum_{t=2}^{T}(Y_{t-1}^2)}.$$

L'erreur-type de ce coefficient est :

$$\mathrm{SE}(\hat{\delta}) = \sigma \Big/ \sqrt{\sum_{t=2}^{T}(Y_{t-1}^2)},$$

où :

$$\hat{\sigma}^2 = \frac{1}{T-1}\sum_{t=2}^{T}(\Delta Y_t - \hat{\delta} \cdot Y_{t-1})^2.$$

Ainsi, la statistique de test de Dickey-Fuller s'écrit :

$$\tau = \frac{\hat{\delta}}{\mathrm{SE}(\hat{\delta})}.$$

11.6 Distribution sous H_0

Sous l'hypothèse nulle $H_0 : \delta = 0$, la distribution de la statistique *tau* n'est pas une Student classique, elle suit la distribution de Dickey-Fuller, tabulée par Dickey et Fuller. Ainsi, le rejet de H_0 consiste à comparer la valeur calculée de *tau* avec le quantile de la table de Dickey-Fuller.

11.7 Décision

$$\text{Si} \quad tau < \mathrm{CV}_\alpha \quad \rightarrow \quad H_0 \text{ rejetée (série stationnaire)},$$
$$\text{Sinon} \quad H_0 \text{ non rejetée (série non-stationnaire)},$$

où :

CV_α est la valeur critique de Dickey-Fuller au seuil α.

12 Dickey-Fuller Augmenté (ADF)

Toutefois, dans la réalité des données, ce terme peut être *autocorrélé*. Pour corriger ce problème, le test de Dickey-Fuller Augmenté (ADF) consiste donc à inclure des décalages de la variable afin d'absorber l'autocorrélation des résidus.

12.1 Modèle

Le modèle ADF s'écrit :

$$\Delta Y_t = \delta \cdot Y_{t-1} + \sum_{i=1}^{p} \beta_i \cdot \Delta Y_{t-i} + \epsilon_t,$$

où :

- Y_t est la valeur de la série chronologique au temps t.
- δ est le coefficient que l'on teste contre 0 (comme dans le cas simple).
- p est le nombre de décalages (lags) inclus afin d'éliminer l'autocorrélation des résidus.
- β_i sont les coefficients des décalages de la différence de la série.
- ϵ_t est un bruit blanc i.i.d.

12.2 Lien avec le Dickey-Fuller simple

Le test de Dickey-Fuller simple est le cas particulier de l'ADF lorsque $p = 0$:

$$\Delta Y_t = \delta \cdot Y_{t-1} + \epsilon_t.$$

12.3 Hypothèses

$$\begin{cases} H_0 : \delta = 0 & (\rho = 1 \quad \rightarrow \quad \text{Série non-stationnaire}) \\ H_1 : \delta < 0 & (\rho < 1 \quad \rightarrow \quad \text{Série stationnaire}) \end{cases}$$

12.4 Statistique de test

De la même manière que dans le cas simple, le test consiste à estimer le modèle par moindres carrés ordinaires, puis à former la statistique :

$$\tau_{ADF} = \frac{\hat{\delta}}{\mathrm{SE}(\hat{\delta})},$$

où :

$$\hat{\delta} \quad \text{est l'estimateur de } \delta$$
$$\text{SE}(\hat{\delta}) \quad \text{est l'erreur-type de } \hat{\delta}.$$

12.5 Distribution sous HH_0

Tout comme le Dickey-Fuller simple, la distribution de la statistique ADF sous l'hypothèse nulle n'est pas Student, mais suit la *distribution de Dickey-Fuller*. Les valeurs critiques sont donc générées par simulation de Monte Carlo et sont tabulées dans la plupart des manuels d'économétrie.

12.6 Décision

$$\text{Si} \quad tau_{ADF} < \text{CV}_\alpha \quad \rightarrow \quad H_0 \text{ rejetée (série stationnaire)},$$
$$\text{Sinon} \quad H_0 \text{ non rejetée (série non-stationnaire)},$$

où :

$$\text{CV}_\alpha \quad \text{est la valeur critique de Dickey-Fuller augmentée au seuil } \alpha.$$