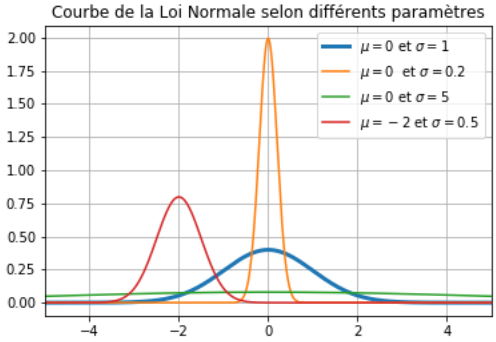
Estimateurs	**Formule**	**Définition**
Moyenne	$\bar{x} = \frac{1}{n}\sum x_i$	Mesure de la tendance centrale d'une série
Variance	$\sigma_x^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$ ** On retire un degré de liberté pour avoir un estimateur non biaisé.*	Mesure de la dispersion d'une série
Écart-type ou *Standart Deviation* (STD)	$Std(X) = \sqrt{Var(X)}$	Mesure de la dispersion indépendamment de l'unité de la série
Covariance	$\sigma_{xy} = \frac{1}{n-1}\sum (x_i - \bar{x})\ (y_i - \bar{y})$	Mesure pour quantifier la relation entre deux séries
Corrélation	$r(X,Y) = \frac{Cov(X,Y)}{Std(X)Std(Y)}$	Mesure normalisée entre 0 et 1 de la covariance (relation entre deux séries)
Erreur-type ou *Standart Error* (SE)	$\sigma_{\bar{x}} = \frac{Std(X)}{\sqrt{n}}$	Mesure de la dispersion des moyennes (issus d'autre tirages) autour de la moyenne de la véritable population

Loi de probabilité

1 – Loi normale

La distribution normale est donnée par :

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



où :

μ est la moyenne
σ est l'écart-type

La distribution Gaussienne ou de la Loi normale centré réduite est noté $Z \sim \mathcal{N}(\mu, \sigma)$.

$$Z = \frac{X_i - \mu_i}{\sigma_i}$$

La loi normale est utile en raison **du théorème de la limite centrale** (TCL) : la somme de variables aléatoires de mêmes paramètres sera approximativement distribuée normalement, quelle que soit leurs distributions (lorsque les effectifs sont supérieurs à 30 en pratique).

2 – Loi du χ^2

La distribution du **Khi-deux** noté χ^2_n est la somme au carré de n loi normales centrés réduite.

$$U = \sum_{i=1}^{k} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$$

La distribution χ^2 est utilisée pour modéliser les erreurs comme la **somme des carrés** ou la **distribution de la variance** d'un échantillon.

3 – Loi de Fisher

La distribution de Fisher noté $F_{n,p}$ est le ratio entre deux lois du χ^2.

$$F_{n,p} = \frac{U/n}{U/p}$$

Cette distribution permet de tester de manière générale **l'égalité des variances** entre deux variables ou si le **ratio entre deux erreurs** est suffisamment élevé.

4 – Loi de Student

La **Loi de Student** est le quotient entre une variable suivant une loi normale centrée réduite et la racine carrée d'une variable distribuée suivant la loi du χ^2.

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

* Lorsque k (degré de liberté) est grand, la loi de Student peut être approchée par la loi **normale centrée réduite**.

Tests statistiques

P-value : c'est la probabilité de trouver pour un modèle une même valeur, ou une valeur plus extrême, sous l'hypothèse nulle (H_0). Elle indique dans quelle mesure (probabilité) les données sont **conformes** à l'hypothèse de test (généralement l'hypothèse nulle).

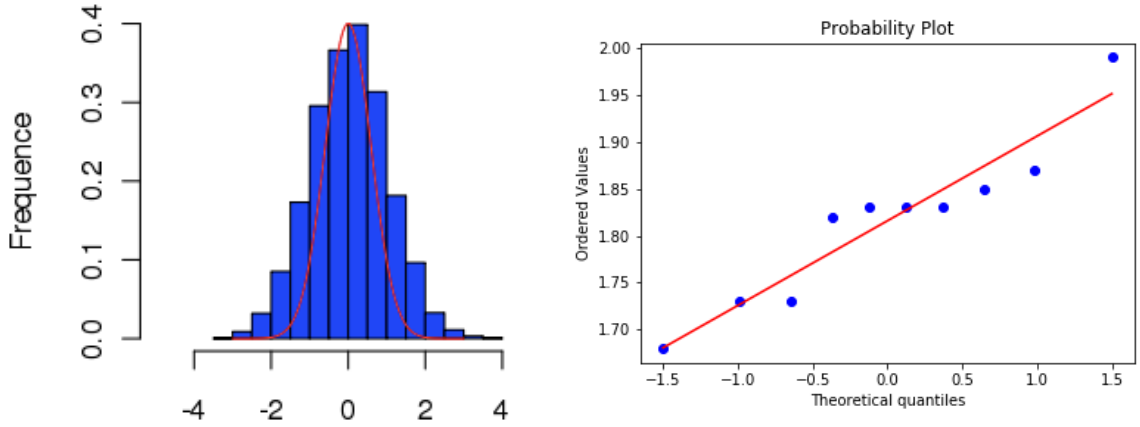
1. Test paramétrique

Les **tests paramétriques** sont des tests qui se basent sur la distribution supposée des données (loi normale, égalité des variances...). Par conséquent, certaines conditions de validité doivent être vérifiées pour que le résultat d'un test paramétrique soit fiable.

2.1. Test de normalité

On peut vérifier si les données suivent une loi normale avec :

1 – Histogramme de la distribution ou Graphique des QQ-plot



2 – Les tests

L'hypothèse nulle (H_0) des tests de normalités est :

- H_0 : « la distribution de la variable suit une loi normale »
- H_1 : « la distribution de la variable ne suit pas une loi normale »

⇨ Si la p-value est **inférieure** à un niveau α choisi (par exemple 0.05), alors l'hypothèse nulle est rejetée (i.e. il est improbable d'obtenir de telles données en supposant qu'elles soient normalement distribuées).

⇨ Si la p-value est **supérieure** au niveau α choisi (par exemple 0.05), alors on ne doit pas rejeter l'hypothèse nulle. La valeur de la p-value alors obtenue ne présuppose qu'il s'agit vraisemblablement d'une **distribution normale**.

Test Kolmogorov-Smirnov (KS) : le test se base sur la fonction de répartition de la loi normale comparé à celle de l'échantillon.

Test d'Agostino : le test se base sur l'ampleur de l'asymétrie de la distribution (selon le Kurtosis).

Test de Shapiro-Wilk : le test est basé sur la corrélation entre les données et le score théoriques de la loi normale.

* Il faut noter qu'il est possible de se rapprocher d'une loi normale en transformant la variable avec une fonction (log ou racine carrée).
* Certains tests sont **robustes** au test de normalité (leurs résultats ne varient pas trop sauf si la distribution est vraiment bimodale).
* Lorsque la taille des échantillons est suffisamment grande (n >30), on peut ignorer le test de normalité grâce au **théorème centrale limite**.

2.2. Test d'égalité des variances : test F

Il s'agit du rapport entre la variance la plus élevée et la variance la plus faible des échantillons.

L'hypothèse nulle (H_0) du test F est :
- H_0 : « les variances des deux groupes sont égales ».
- H_1 : « les variances des deux groupes ne sont pas égales ».

$$F = \frac{S^2_{max}}{S^2_{min}}$$

avec : $F_{n,p}$ où n est $(N_{max} - 1)$ et p est $(N_{min} - 1)$ degré de liberté.

⇨ Si le test est significatif (la p-value est inférieure au risque α ou que $F_{observé} > F_{théorique}$), l'hypothèse nulle est rejetée et on peut conclure que les variances sont **significativement différentes** (ex : p-value est égale à 0.001).

⇨ Si le test n'est pas significatif (la p-value est supérieure au risque α ou que le $F_{observé} < F_{théorique}$), alors on ne rejette pas l'hypothèse nulle et on peut supposer que les **variances sont égales** (ex : p-value est égale à 0.96).

2.3. Test de Student

Avant d'utiliser le test de Student il faut vérifier :

1 – Dans le cas du test de **Student** pour un **échantillon unique** :
- Si les données suivent la **loi normale**.

2 – Dans le cas du test de **Student indépendant** entre **deux échantillons** :
- Si les deux groupes d'échantillons suivent une **loi normale** ;
- Si les **variances** des deux groupes sont **égales** ou pas.

3 – Pour le test de **Student apparié** (il existe un lien entre les deux échantillons) :
- Si la différence entre les deux échantillons suit une **loi normale**.

1 | Test de Student à échantillon unique

Le test de Student pour un échantillon unique compare la **moyenne observée** ($\bar{X}$) d'un échantillon avec une **moyenne théorique** (μ) en prenant en compte le nombre d'observations (N) et la variance de l'échantillon (s^2).

Les hypothèses du test t sont :
- H_0 : « la moyennes sont égales à la moyenne théorique ».
- H_1 : « la moyennes est différente de la moyenne théorique ».

$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{N}}}$$

avec : t suit une loi de Student à (n-1) degrés de liberté.

```
scipy.stats.ttest_1samp(x, mu)
```

Exemple : test de Student pour tester une moyenne théorique (1,75).

```
x = [1.83, 1.83, 1.73, 1.82, 1.83,
     1.73, 1.99, 1.85, 1.68, 1.87]

xbar = np.mean(x)
s2 = np.var(x, ddof=1)
n = len(x)

t = (xbar - 1.75) / (np.sqrt(s2 / n))
print(t)
```

```
2.3968766311585883
```

Ce test est applicable seulement si les valeurs de X **suivent une loi normale**

2 | Test de Student pour deux échantillons

Le test de Student pour deux **échantillons indépendants** compare si les moyennes observées pour deux échantillons sont équivalentes. Plusieurs tests en découlent selon la forme des échantillons :

Les hypothèses du test t sont :
- H_0 : « les moyennes deux échantillons sont égales ».
- H_1 : « les moyennes des deux échantillons sont différentes ».

⇨ Si les **variances ne sont pas égales** (test de Welch's), généralement dû à des effectifs différents :

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Pour déterminer le nombre de degré de liberté :

$$\nu = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2 \cdot (N_1 - 1)} + \frac{s_2^4}{N_2^2 \cdot (N_2 - 1)}}$$

```
scipy.stats.ttest_ind(x1,x2, equal_var=False)
```

⇨ Si les variances sont **égales** :

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

Avec la variance commune calculée par :

$$s = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}}$$

avec $(n_1 + n_2)$ - 2 degrés de liberté.

* Lorsque les effectifs sont très déséquilibrés il est préférable d'utiliser la formule avec l'hypothèse d'inégalité des variances, même si les variances sont égales.

```
scipy.stats.ttest_ind(x1,x2, equal_var=True)
```

Ces tests sont applicables seulement si les valeurs de X **suivent une loi normale**

3 | Test de Student pour deux échantillons appariés

Le test de **Student apparié** permet de comparer la moyenne de deux séries de valeurs **ayant un lien**. Ce lien existe lorsque les séries répondent à une de ces deux conditions : (1) les individus de chaque paire se ressemblent le plus possible ou (2) appartiennent à une même entité statistique.

Pour comparer les moyennes de deux séries appariées, on calcule tout d'abord la **différence des deux mesures** pour chaque paire. Ensuite, on teste si cette nouvelle série ($\bar{d}$) est significativement différente de 0, comme pour un test à échantillon unique.

Les hypothèses du test t sont :
- H_0 : « la différence moyenne entre les deux échantillons est égales ».
- H_1 : « la différence moyenne entre les deux échantillons est différentes ».

$$t = \frac{\bar{d}}{\sqrt{\frac{s^2}{N}}}$$

Ce test est applicable seulement si la différence **suit une loi normale.**

2.4. ANOVA

L'analyse de la variance (ANOVA) est un test statistique permettant de déterminer si les **moyennes de plusieurs groupes** sont égales ou non, et donc de généraliser le test t à plus de deux groupes avec pour degrés de liberté $F(p-1, n-p)$.

Les hypothèses du test F sont :
- H_0 : « les moyennes des échantillons sont équivalentes ».
- H_1 : « au moins deux moyennes sont différentes ».

$$F = \frac{\frac{SCE_{\text{expliquée}}}{p-1}}{\frac{SCE_{\text{residu}}}{n-p}}$$

On calcule la **somme des carrés des écarts expliquée** :

$$SCE_{\text{expliquée}} = \sum_{i=1}^{p} n_i \, (\overline{y_i} - \overline{y})^2$$

Une valeur de SCE élevée indique que les **moyennes entre les échantillons** sont très **différentes** (variabilité inter-classe). Il s'agit d'un calcul de variance entre les différents groupes.

On calcule la **somme des carrés des écarts résiduelle** :

$$SCR_{\text{residu}} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(y_i^j - \overline{y_i}\right)^2$$

Une valeur de SCR élevée par rapport à la somme des carrés totaux indique que la variabilité au sein de chaque groupe est très élevée (variabilité intra-classe). Par conséquent, la **différenciation** entre chaque groupe est **faible (SCE).**

* Il faut noter que pour un test entre deux échantillon, $F = t^2$ où t est la statistique de Student.
* L'ANOVA est très robuste aux hypothèses de normalités et d'homoscédasticité.

`scipy.stats.f_oneway(a,b,c)`

Exemple : test de Fisher pour 3 échantillons.

```
import numpy as np

a = np.array([20.1, 19.8, 21.3, 20.7])
b = np.array([22.6,24.1,23.8,22.5,23.4,24.5,22.9])
c = np.array([31.2,31.6,31.0,32.1,31.4])

# Calcul de la moyenne globale
arg = [i for i in [a,b,c]]
x = np.concatenate(arg)
mean = np.mean(x)

SCE = list()
SCR = list()
SCT = list()

# Calcul de la variance inter-classe
for i in [a,b,c]:
    SCE.append(len(i) * ((np.mean(i) - mean)**2))
print(f"SCE = {sum(SCE)}")

# Calcul de la variance intra-classe
for i in [a,b,c]:
    mean_ = np.mean(i)
    for j in i :
        SCR.append((j - mean_)**2)
print(f"SCR = {sum(SCR)}")

# Calcul de la variance totale
SCT.append(np.var(x, ddof=1) * (len(x) - 1))
print(f"SCT = {SCT[0]}")

# Calcul des carrés moyens résiduels et expliqués
CME = sum(SCE) / (3 - 1)
CMR = sum(SCR) / (len(x)-3)

print(f'F = {CME / CMR}')
```

```
SCE = 307.9179999999999
SCR = 5.5995000000000035
SCT = 313.5174999999999
F = 357.43673542280527
```

L'exemple montre que la statistique de Fisher est très élevée. Ainsi, au seuil α *de 5%, on peut facilement rejeter l'hypothèse nulle* H_0 et conclure que au ***moins deux moyennes sont différentes avec un risque d'erreur de 5%.***

Ce test est applicable seulement si **(1)** les échantillons **suivent une loi normale** et **(2)** la **variance des groupes sont identiques** (homoscédasticité).

2.5. Test de χ^2 d'indépendance

Ce test permet tester l'indépendance entre deux variables. Il s'agit de calculer la distance au carré entre les données observés et les données théoriques représentées par un **tableau de contingence** (cas d'une indépendance des données).

Si la distance est trop élevée entre les deux tableaux, on peut considérer que les données sont trop différentes de l'hypothèse H_0 (hypothèse d'indépendances des deux séries) et donc conclure sur la **dépendance** des données.

Les hypothèses du test de χ^2 d'indépendance sont :
- H_0 : « les variables sont indépendantes ».
- H_1 : « les variables sont dépendantes ».

$$E_{théorique} = \frac{n_{i+}\ n_{+j}}{n}$$

où n_{i+} est l'effectif total de la ligne i
où n_{+j} est l'effectif total de la colonne j

$$T\ = \sum_{i,j} \frac{\left(E_{réel} - E_{théorique}\right)^2}{E_{théorique}}$$

T suit une loi du χ^2 à (ligne – 1) (colonne – 1) degrés de liberté.

Si la p-value (probabilité que $\chi^2_{\text{théorique}}$ dépasse le $\chi^2_{\text{calculé}}$ sous l'hypothèse H_0) est :
- supérieur au risque α , on **ne rejette pas** l'hypothèse H_0 (on accepte l'indépendance).
- inférieur au risque α, on **rejette** l'hypothèse H_0 (on accepte la dépendance).

Exemple :

```python
import numpy as np
import pandas as pd
import scipy.stats as stats

X = np.array(['A'] * 290 + ['B'] * 285)
Y = np.array([1] * 50 + [2] * 70 + [3] * 110 + [4] * 60 +
             [1] * 60 + [2] * 75 + [3] * 100 + [4] * 50)

crosstab = pd.crosstab(X, Y)

print("Observed table:")
print("---------------")
print(crosstab)
print("")

t = len(Y[Y == 1]) * len(X[X =='A']) / len(Y)
print(f"i1,j1 = {t}\n")
chi2, pval, dof, expected = stats.chi2_contingency(crosstab)

print("Expected table:")
print("---------------")
print(expected)
print("")

print("Statistics:")
print("-----------")
print("Chi2 = %f, pval = %f ddof= %f" % (chi2, pval, dof))
```

```
Observed table:
---------------
col_0   1   2    3   4
row_0
A      50  70  110  60
B      60  75  100  50

i1,j1 = 55.47826086956522

Expected table:
---------------
[[ 55.47826087  73.13043478 105.91304348  55.47826087]
 [ 54.52173913  71.86956522 104.08695652  54.52173913]]

Statistics:
-----------
Chi2 = 2.423491, pval = 0.489277 ddof= 3.000000
```

2. Test non paramétriques

3.1. Distance Euclidienne (L_2)

Distance entre deux vecteurs à n dimensions, également appelé **Norme 2** :

$$|| x - y ||_2 = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$

3.2. Distance de Manhattan (L_1)

Distance donnée par la somme des valeurs absolues, également appelée **Norme 1** :

$$|x - y|_1 = \sum_{i=0}^{n} |x_i - y_i|$$

3.2. Distance de Mahalanobis

Mesure la distance entre plusieurs variables selon la corrélation. Elle accorde un poids moins important aux composantes les plus dispersées.

$$D_M(x) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}$$