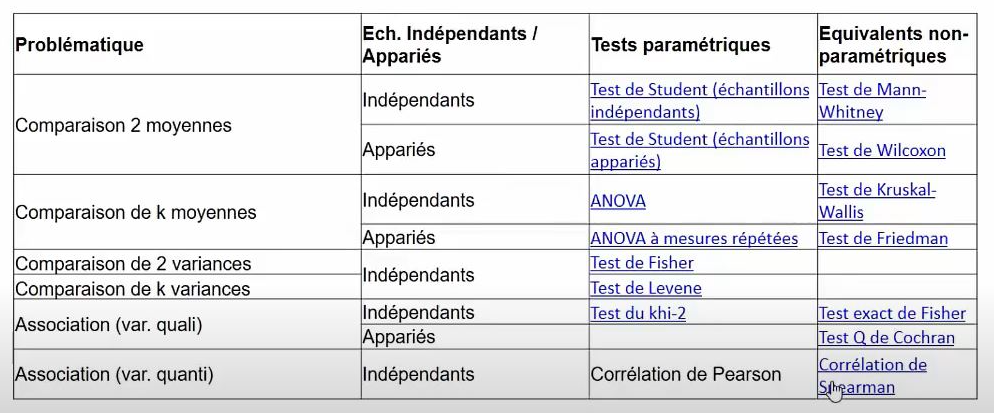
Choisir ces tests statistiques en fonction de notre problématique

1. **Question** : à formuler (rep oui ou non)
2. **Les hypothèses**:
   1. **Ho hypothèse nulle** (généralement ==)
   2. **Ha hypothèse alternative** (généralement diff.)
3. **Observations** :
   1. **Comparaison ou Association ?** 
      1. **Comparaison** : qu’est-ce qu’on compare :
         1. 2 moyennes
         2. K moyennes
         3. 2 variances
         4. K variances
         5. Autres... (écart-type…)
      2. **Association** : va dépendre du type de variable
         1. Variables qualitatives
         2. Variables quantitatives
   2. **Echantillons indépendants ou appariés ?**
      1. **Indépendants** = deux ou plusieurs populations distinctes.
         1. Ex : Bananes traitées, bananes non traitées
         2. Homme / femme
         3. Groupe témoin / groupe exposé
      2. **Appariés** = une seule population d’individus
         1. Ex : Mesures sur des patients avec ttt avant/après
         2. Dans le temps : suivi de personnes, d’une entreprise a différentes dates
         3. Tx de sucre de bananes groupe A avec différents engrais
   3. **Tests paramétriques et non paramétriques :**
      1. **T paramétriques** :
         1. Puissant
         2. Dde des conditions de normalité vérifiées au niv de la distri (hypothèse de normalité vérifiée)
         3. Souvent apte à rejeter la Ho si les conditions le permettent
      2. **T non paramétriques** :
         1. Robuste
         2. N’assument pas de distri dans les populations (calculé la plupart du temps par rapport à un groupe de D)
         3. Sont fiable sur un grand nb de cas
4. **Choix du test** les plus courant :

**Non para :**

\*\*coeff de Spearman\*\* : CATEGORIEL (non para) également appelé rhô ou r(s), évalue l'association entre deux variables mesurées dans une échelle ordinale . Il assume des valeurs se situant dans l'intervalle qui va de - 1 à +1 (Positive correlations imply x augmente y aussi. Negative correlations imply x augmente, y diminue.) permet de détecter des tendances monotones. Lorsque la tendance est affine, il se comporte de façon similaire au coefficient de Pearson

conditions du test : test non para donc présume que l'échantillon ne suit pas une loi normale. L'autre condition de validité pour le calcul d'un coefficient de corrélation ou l'estimation d'une régression linéaire, est l'existence d'une variance non-nulle sur chacune des deux variables, sous peine de division par zéro ( c a d ps ls mm d partt)

Alternative : Coeff de Pearson (para) continue

Code : (autre manière d’afficher la pvalue) MAIS REGARDER LES EXPOSANTS NEGATIFFFFFFS

from scipy.stats import spearmanr  
alpha = 0.05

corr, p = scipy.stats.spearmanr(ages\_montant1.age, ages\_montant1.price)  
print(f'stat=%.3f, p=%.3f, ' % (corr, p))  
if p < alpha:  
 print('H0 rejetée')  
else:  
 print('H0 non rejetée')

2ème méthode :  
resultat = scipy.stats.spearmanr(ages\_montant1.age, ages\_montant1.price)  
print("stat de test", round(resultat[0],3))  
print("p-value", resultat[1])

**PARA :**

\*\*Coeff de Pearson\*\* : VAR CONTINU (para) exprime l'intensité et le sens (positif ou négatif) de la relation linéaire entre deux variables quantitatives . Il assume des valeurs se situant dans l'intervalle qui va de - 1 (nég) à +1 (,0 = pas de corr. Plus ce coefficient est proche de -1 ou +1, plus l'association entre les deux variables est forte, jusqu'à être parfaite.

Condition : est l'existence d'une variance non-nulle sur chacune des deux variables, sous peine de division par zéro, et comme paramétrique échantillon doit suivre une loi normale

(Le coefficient de corrélation linéaire nous aide à juger de l’existence d’une relation **linéaire** entre deux variables c’est-à-dire lorsque l’on peut tracer une ligne droite dans le nuage de points. Il n’est donc pas adapté lorsque les relations ne sont pas linéaires. Attention : corrélations « absurdes ». une corrélation ne constitue pas une preuve de relation de cause à effet (causalité)

\*\* ANOVA\*\* : L'analyse de la variance (**ANOVA**) Le **test** F **est utilisé** dans le processus d'**ANOVA pour** tester la différence entre les moyennes ou l'égalité de la **variance**. L'**ANOVA** sépare la variabilité intra-échantillon de la variabilité inter-échantillons. Le **test** F **est** le rapport de l'erreur quadratique moyenne de ces deux échantillons

Conditions : doit suivre une loi normale et égalité des variances selon comment on l’utilise

Khi-2 :

H0 : variables sont indépendantes

H1 : variables non indépendantes

Conditions : para donc suit normalité, deux groupes ou deux mesures (var qualix2)

Faire un tableau de contingence classique puis par fréquence %

indique si les fréquences d’observation d’une catégorie dépendent des modalités de l’autre catégorie 🡪 va indiquer les fortes corrélations pour certaines valeurs de nos deux variables

-Calcul de la stat khi2 (statistic)

-Calcul du nombre de degrés de liberté k (dl)

-Calcul de la loi à densité suivie par une loi khi2 à k degrés de liberté + Confrontation de la stat khi2 calculée avec la loi à densité obtenue (🡪 pvalue = indique quelle probabilité (matérialisée par l’aire sous la courbe de la loi à densité) représente les valeurs encore plus éloignées que celle obtenue)

V-Cramer : mesure de la force de l'association entre deux [variables nominales](https://www.statology.org/levels-of-measurement-nominal-ordinal-interval-and-ratio/) .

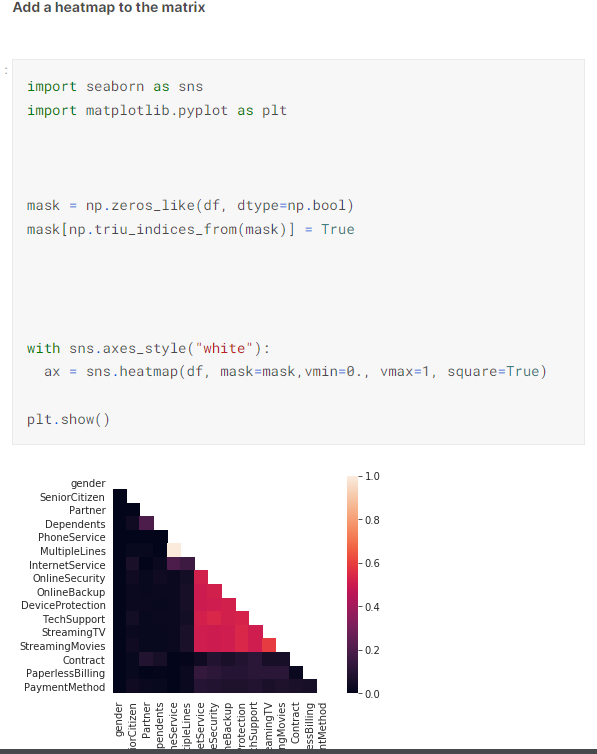
Il varie de 0 à 1 où :

* **0** indique aucune association entre les deux variables.
* **1** indique une forte association entre les deux variables.

se base sur le coeff de corr du k2 (‘x2’) pour trouver quelle corrélation exactement il y a entre les variables : a faire après un k2 si on veut etre plus précis

X2=20.20 🡪 (résultat stat du k2)  
n = np.sum(tab\_contingence) 🡪 somme de toutes nos valeurs du tableau de contingence  
minDim = min(tab\_contingence.shape)-1 🡪 prend la val minimum entre les colonnes et les lignes de notre tableau de contingence

Possibilité de faire un heatmap de la matrice :



**Test de normalité :**

H0 : échantillon suit une loi normale

H1 : ne suit pas une loi normale

Ressources intéressantes : http://eric.univ-lyon2.fr/~ricco/cours/cours/Test\_Normalite.pdf

Kolmogorov-Smirnov

Test non paramétrique MAIS ATTENTION

Avec le paramètre ‘norm’ on va comparer nos échantillons à la loi normal avec ‘norm’

Attention : Avec le ks\_2sample c’est pour comparer deux échantillons s’ils proviennent de la mm distri ( voir fiche)

Test de D'Agostino et Pearson - normaltest

scipy.stats.normaltest()

 Il est basé sur le test de D'Agostino et Pearson [[1]](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html#r7bf2e556f491-1) , [[2]](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html#r7bf2e556f491-2) qui combine skew et kurtosis (mesures d’asymétrie et d’aplatissement) pour produire un test omnibus de normalité pour un échantillon de taille moyenne et grande

Sortie : statistic du test = k2 et pvalue

test de Shapiro Wilk : from scipy.stats import shapiro

Attention jusqu’à 50 valeurs max !

-Résultat stat : basé sur la corrélation, + il est proche de 1 mieux il montre que les D correspondent à la distri normale  
-pvalue

**Test d’égalité des variances :**

**H0** = égalité des variances

**H1** : inégalité des variances

Existe test F, Test de Levene ou Bartlett, Fligner (non param).... Plus ou moins sensible si les échantillons ne suivent pas une loi normale

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.fligner.html#scipy.stats.fligner>

bartlett : **from** scipy.stats **import** bartlett

test para d’égalité de k variances dans des échantillons normaux :

**condition** : suit une loi normale

[levene](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.levene.html#scipy.stats.levene) : **from** scipy.stats **import** levene

test para mais robuste donc a priori pas obligé suivent une loi normale (3 variantes qui diffèrent par la mesure de la tendance centrale   
Le test de Levene est une alternative au test de Bartlett [bartlett](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bartlett.html" \l "scipy.stats.bartlett" \o "scipy.stats.bartlett) dans le cas où il existe des écarts significatifs par rapport à la normalité.

**Condition**: mettre bien la bonne variante ! car loi normale ou pas mais préciser !!

Trois variantes du test de Levene sont possibles. Les possibilités et leurs utilisations recommandées sont :

* 'median' : Recommandé pour les distributions asymétriques (non normales)>
* 'mean' : Recommandé pour les distributions symétriques à queue modérée.
* 'trimmed' : Recommandé pour les distributions à queue lourde.

Fligner **from** scipy.stats **import** fligner

Le test de Fligner-Killeen est sans distribution lorsque les populations sont identiques

Tableaux d'échantillons de données n'a pas besoin d'être de la même longueur.

Comme pour le test de Levene, il existe trois variantes du test de Fligner qui diffèrent par la mesure de la tendance centrale utilisée dans le test

**Test d’égalité des moyennes :**

**from** scipy **import** stats

**H0** = égalité des moyennes

**H1** : inégalité des moyennes

t-test :

Calculez le test T pour les moyennes de deux échantillons indépendants.

Il s'agit d'un test pour l'hypothèse nulle selon laquelle 2 échantillons indépendants ont des valeurs moyennes (attendues) identiques.

CONDITIONS : Ce test suppose que les populations suivent une loi normale et ont des variances identiques par défaut (paramètre equal\_var=False ou True selon cas)

ATTENTION : [ttest\_ind](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html#scipy.stats.ttest_ind) sous-estime p pour des variances inégales

Le test t quantifie la différence entre les moyennes arithmétiques des deux échantillons. La valeur de p quantifie la probabilité d'observer une ou plusieurs valeurs extrêmes en supposant que l'hypothèse nulle, selon laquelle les échantillons sont tirés de populations ayant les mêmes moyennes de population, est vraie.

Une valeur de p > alpha = peu probable que notre observation se soit produite par hasard = non rejet de l’H0  
Une valeur de p < alpha = nous avons des preuves contre l'hypothèse nulle de moyennes de population égales.

Par défaut, la valeur de p est déterminée en comparant la statistique t des données observées à une distribution t théorique. Quand , où1 < permutations < binom(n, k)

* K est le nombre d'observations dans *a* ,
* N est le nombre total d'observations dans *a* et *b* , et
* binom(n, k) est le coefficient binomial ( n choisir k),

les données sont regroupées (concaténées), attribuées au hasard au groupe *a* ou *b* , et la statistique t est calculée. Ce processus est effectué à plusieurs reprises ( temps de *permutation* ), générant une distribution de la statistique t sous l'hypothèse nulle, et la statistique t des données observées est comparée à cette distribution pour déterminer la valeur p.

A savoir : L'utilisation de l'ajustement est communément appelée test t ajusté. Parfois appelé test t de Yuen, il s'agit d'une extension du test t de Welch, la différence étant l'utilisation de moyennes winsorisées dans le calcul de la variance et la taille de l'échantillon réduit dans le calcul de la statistique. Le découpage est recommandé si la distribution sous-jacente est à longue queue ou contaminée par des valeurs aberrantes

stats**.**ttest\_ind**(**data1**,** data2**)**

stats**.**ttest\_ind**(**data1**,** data2**,** equal\_var**=False)**

On doit vérif la normalité, l’égalité des variances en amont

**VOCABULAIRE :**

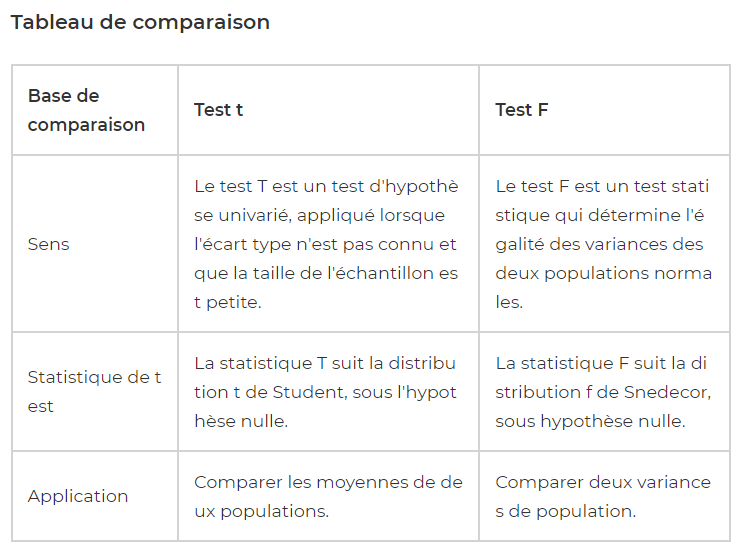
ERREUR MOYENNE QUADRATIQUE

Quadratique : élevé au carré. (pour pas que résultat soit nul)

COMPARAISON d’estimateurs : Il s’agit d’une fonction de risque, correspondant à la valeur attendue de la perte d’erreur au carré (non null) , cette moyenne d’erreur au carré : permet donc de trancher dans une situation où il existe un estimateur sans biais et un autre biaisé mais de variance plus petite (valeur estimée et la valeur vraie)

Test omnibus :

**Les tests Omnibus** sont des [tests statistique](https://stringfixer.com/fr/Statistical_test)s . Ils testent si la variance expliquée (attendue) dans un ensemble de données est [significativement](https://stringfixer.com/fr/Statistical_significance) supérieure à la [variance](https://stringfixer.com/fr/Variance) inexpliquée (théorique). Un exemple est le [test F](https://stringfixer.com/fr/F-test) dans l' [analyse de la variance](https://stringfixer.com/fr/Analysis_of_variance) ou le [test du](https://stringfixer.com/fr/F-test)[Chi-carré](https://stringfixer.com/fr/Chi-squared_test) .. Il peut y avoir des effets significatifs légitimes dans un modèle même si le test omnibus n'est pas significatif

Test T, test F…. :

### Définition du test t

Un test t est une forme du test d'hypothèse statistique, basé sur la statistique t et la distribution t de Student pour rechercher la valeur p (probabilité) pouvant être utilisée pour accepter ou rejeter l'hypothèse nulle..

Le test T analyse si les moyennes de deux ensembles de données sont très différentes l'une de l'autre, c'est-à-dire si la moyenne de la population est égale ou différente de la moyenne standard. Il peut également être utilisé pour déterminer si la droite de régression a une pente différente de zéro

### Définition du test F

Le test F est décrit comme un type de test d'hypothèse basé sur la distribution f de Snedecor, sous l'hypothèse nulle. Le test est effectué quand on ne sait pas si les deux populations ont la même variance.

Le test F peut également être utilisé pour vérifier si les données sont conformes à un modèle de régression, acquis au moyen d'une analyse par les moindres carrés. Lorsqu'il existe une analyse de régression linéaire multiple, elle examine la validité globale du modèle ou détermine si l'une des variables indépendantes entretient une relation linéaire avec la variable dépendante

## Principales différences entre le test T et le test F

La différence entre t-test et f-test peut être clairement établie pour les motifs suivants:

1. Le test t est un test d'hypothèse univarié appliqué lorsque l'écart type n'est pas connu et que la taille de l'échantillon est petite. D'autre part, un test statistique, qui détermine l'égalité des variances des deux ensembles de données normaux, est appelé test f..
2. Le test t est basé sur la statistique t qui suit la distribution t de Student, sous l'hypothèse nulle. Inversement, la base du test f est que la statistique F suit la distribution f de Snedecor, sous l'hypothèse nulle.
3. Le test t est utilisé pour comparer les moyennes de deux populations. En revanche, le test f est utilisé pour comparer deux variances de population.