**LES TESTS STATISTIQUES**

**Test t de Student** : porte sur une variable qualitative (ordinale ou nominale) qui contient 2 modalités. Il s’agit de comparer la moyenne d’une variable continue (une variable quantitative) entre 2 groupes / échantillons.

*Définition : Les échantillons sont indépendants si on ne retrouve pas les mêmes individus dans chacun des échantillons.*

**Test Anova** : même principe que le t-test (on compare la moyenne de plus de 2 échantillons. C’est une généralisation du t-test. L’idée repose sur la comparaison de la variance entre les groupes avec la variance aléatoire au sein de chaque groupe. Si la variance entre les groupes est significativement plus grande qua la variance aléatoire, alors les moyennes sont probablement également différentes.

*Définition : on compare la variation intragroupe (l’ampleur de la variation au sein d’un groupe) avec la variation intergroupe (mesure les variations de moyennes entre les différents groupes)*

* *Si F est supèrieur à 1, cela signifie que la variation intergroupe est supérieure à la variation intragroupe. Plus F s’éloigne de 1, plus les écarts sont importants. Un F égal à 2 signifie qu’il y a 2 fois plus d’écarts intergroupes que d’écarts intragroupes.*

Si la significativité est < à 0,05 donc bien un écart de moyenne entre les différents groupes.

**Formaliser le problème de test**

1. **Les hypothèses d’un test :**

H0 = hypothèse nulle (celle que nous ne souhaitons pas rejeter)

H1 = hypothèse alternative (nous donne les conditions pour lesquelles on rejette H0

1. Région critique : indique quand on rejette l’hypothèse nulle, elle est basée sur une statistique de test. Pour la déterminer il faut connaitre la loi de la statistique de test sous H0.
2. Risques liés à la décision : si on rejette H0 alors que cette hypothèse est vraie, c’est une erreur de première espèce, puis si on rejette H1 alors que cette hypothèse est vraie, c’est une erreur de seconde espèce. Les valeurs couramment employées sont 5%, 1%, et 0,5%.

Une image contenant texte, Police, ligne, capture d’écran

Description générée automatiquement

La bonne configuration se trouve en vert.

**p-valeur :** usuellement on travaille avec la p-valeur du test : la plus petite valeur du niveau de test conduisant au rejet de H0. C’est cette valeur que l’on trouve dans les logiciels statistiques. On construit cette p-valeur de manière à avoir équivalence entre le rejet de H0 et le fait que la p-valeur soit inférieur au niveau de test : {Rejet de H0 au niveau de test α}⇔{p-valeur<α}

**RAPPORT STATISTIQUES**

3 ETAPES DE REDACTION

1. Faire un plan
2. Rédiger une première version
3. Finaliser le texte (correction, remaniement, …)

1/ Il faut commencer par la synthèse des résultats et les recommandations

* **Lois de proximité :**

1. Lois de proximité géographique : lecteur sera plus intéressé par des données dans sa ville ou son pays que dans un pays qui se trouve à l’autre bout du monde
2. Loi de proximité temporelle : futur immédiat plus intéressant que futur lointain
3. Loi de proximité professionnelle : information qui touche le lecteur telles que des données sur l’emploi, les impôts, etc.

* **Notion d’angle : divers points de vue selon lesquels un sujet peut être abordé**
* **Aller du particulier au général**

**TEST STATISTIQUES**

**CORRELATION**

La corrélation est un outil statistique utilisé pour le Machine Learning afin d’identifier des relations de dépendance entre plusieurs variables. Il existe plusieurs types de corrélations. Découvrez ci-dessous plus de détails sur les corrélations de Pearson et de Spearman.

Pour l’analyse de données, un [Data Scientist](https://datascientest.com/metier-data-scientist) dispose de [plusieurs outils statistiques](https://datascientest.com/les-outils-dans-le-quotidien-dun-data-scientist). Parmi ces outils, il peut utiliser la corrélation. Il s’agit d’une mesure statistique particulièrement utile qui permet d’étudier la relation entre deux variables à partir du calcul d’un coefficient de corrélation. La corrélation correspond à la force (indiquée par la valeur absolue du coefficient) ainsi qu’à la direction (indiquée par la signe du coefficient) de **la relation entre ces variables**. La direction peut être soit positive (lorsqu’une x augmente, y augmente aussi) ou négative (lorsque x augmente, y diminue ou l’inverse).  Il existe plusieurs types de corrélations. Parmi ces corrélations, il y en a deux qui sont particulièrement utilisées : la corrélation de Pearson et la corrélation de Spearman.

**TEST DE CORRELATIONS PEARSON ET SPEARMAN**

Les corrélations de Spearman et de Pearson sont **deux mesures de corrélation différentes qui s’appliquent dans des situations spécifiques**. La corrélation de Spearman utilise le rang des données pour mesurer la monotonie entre des variables ordinales ou continues. La corrélation de Pearson quant à elle détecte des relations linéaires entre des variables quantitatives avec des données suivant une distribution normale. Dans le cas d’un problème de [Machine Learning](https://datascientest.com/machine-learning-tout-savoir), il est souvent question d’utiliser des matrices de corrélation constituées des coefficients de corrélations entre l’ensemble des variables d’un jeu de données. La notion de corrélation est donc importante pour le Machine Learning.

**LA MONOTONIE ENTRE 2 VARIABLES :**

Dans une **relation monotone, les variables ont tendance à se déplacer dans la même direction relative, mais pas forcément à une vitesse constante.** Dans **une relation linéaire, les variables se déplacent dans la même direction, à une vitesse constante.** Le diagramme 5 montre que les variables augmentent simultanément, mais pas à la même vitesse. Cette relation est monotone, mais pas linéaire. Le coefficient de corrélation de Pearson pour ces données est 0,843, mais celui de la corrélation de Spearman est plus élevé, 0,948.

Les relations linéaires sont également monotones. Par exemple, la relation indiquée sur le diagramme 1 est à la fois monotone et linéaire.

**COEFFICIENT DE PEARSON :**

Avant de calculer le coefficient de Pearson, il faut s’assurer que les données vérifient les hypothèses suivantes :

1. L’échantillon de données est aléatoire (représentatif de la population)
2. Les variables sont quantitatives (continues)
3. Les données sont associées par paires (on associe à chaque valeur x, une valeur y)
4. Les observations sont indépendantes
5. Les données sont distribuées normalement
6. Il existe une relation linéaire entre les variables
7. Aucun outlier n’est présent dans les données

**COEFFICIENT DE CORRELATION DES RANGS DE SPEARMAN**

Avant de calculer le coefficient de Spearman, il faut s’assurer que les données vérifient les hypothèses suivantes :

1. L’échantillon de données est aléatoire
2. La relation entre les variables est monotone
3. Les données sont associées par paires
4. Les observations sont indépendantes
5. Il existe une relation de monotonie entre les variables
6. Les variables sont ordinales ou continues.

La corrélation de rang de Spearman examine la relation entre deux variables. La corrélation de rang de Spearman est le [pendant non paramétrique](https://datatab.fr/tutorial/parametric-and-non-parametric-tests) de la [corrélation de Pearson](https://datatab.fr/tutorial/pearson-correlation). Par conséquent, la distribution normale des données n'est pas requise pour la corrélation de Spearman.

L’interprétation du coefficient de Spearman rs varie selon les valeurs obtenues :

* Si rs est proche de 1, alors il y a une relation monotone positive entre les variables.
* Si rs est proche de 0 alors il n’y a aucune relation monotone entre les variables
* Si rs est proche de -1 alors, il y a une relation monotone négative entre les variables.

Un exemple d’application de la corrélation de Spearman serait l’étude de la relation entre les préférences d’un consommateur et le prix du produit.

**Exemple :**

**Nous souhaitons mettre en relation le délai de jeu de chaque joueur et leur âge**

RANKS reaction time

RANKS reaction time

**AGE**

**TIME**

**d²**

0

1

1

1

1

4

0

0

**Calculer d**

1-1= 0

2-3= -1

3-2= 1

4-5= -1

5-6= -1

7-4= 2

7-7= 0

8-8 = 0

1

2

3

4

5

6

7

8

**12**

14

25

20

35

45

30

60

95

1

3

2

5

6

4

7

8

**15**

**17**

**18**

**20**

**21**

**22**

**26**

8

Nous attribuons les différents rangs à la variables âge selon l’ordre croissant (14 reçoit le rang 1 car il s’agit de la plus petite valeur dans la colonne âge, 20 est la seconde plus petite valeur dans la colonne âge donc il reçoit le rang 2, etc

**Pour le calculer avec le coeff de rang de spearman (coefficient non paramétrique de Pearson) nous devons :**

1. **Ranger du plus petit au plus grand (ordre croissant)**
2. **Donner un numéro de rang à chaque valeur en commençant le rang 1 par la plus petite valeur : dans notre cas 12 a le rang 1, 15 a le rang 2, 17 a le rang 3, 18 a le rang 4, etc.**
3. **On attribue chaque rang aux valeurs de l’âge en respectant l’ordre du plus petit au plus grand : donc 14 qui est la plus petite valeur aura le rang 1 qui avait été attribué à la valeur ‘time’ la plus petite (12) etc**
4. **Visualisons les variables sur un scatter plot (graphique nuage de points) puis comparons sur un second scatter plot où nous formons des rangs des variables ‘âge’ et ‘time’.**

* **A travers cette transformation, nous avons maintenant distribué les données plus uniformément.**

1. **Pour calculer la corrélation de Spearman, nous calculons simplement la corrélation de Pearson à partir des rangs**

* **Donc la corrélation de Spearman est égale à la corrélation de Pearson, sauf que les rangs sont utilisés à la place des valeurs de brutes**

1. **Calculer d et d² pour obtenir le coeff de spearman rs = 1 – (6 x de d²) / 8 x (8² - 1 ) = 1- 48/504 = 0.90**
2. **Comme le coefficient de Pearson r, le coeff de Spearman rs varie également entre -1 et 1**
3. **Nous pouvons déterminer maintenant 2 choses : - la force de la corrélation et la direction qu’elle prend**
4. **FORCE DE CORRELATION : La force de la corrélation peut être lue dans un tableau :**

|  |  |
| --- | --- |
| **Valeur du coefficient de corrélation r** | **Force de la correlation** |
| **0.0 < 0.1** | **Pas de correlation** |
| **0.1 < 0.3** | **Faible correlation** |
| **0.3 < 0.5** | **Correlation modérée** |
| **0.5 < 0.7** | **Forte correlation** |
| **0.7 < 1** | **Trés forte corrélation** |

1. **DIRECTION DE LA CORRELATION**

**entre -1 et 0 : corrélation négative => relation négative entre les variables**

**entre 0 et 1 : corrélation positive => relation positive entre les variables**

1. **EFFECTUER UN test-t de student seulement si l’échantillon est suffisamment grand !**

Si vos données ne suivent pas une distribution normale, certains praticiens vous suggéreront un test non paramétrique (non basé sur l’hypothèse de normalité). D'après son expérience, l'auteur de cet article dirait que si vous avez des données non normales, vous pouvez effectivement considérer le test non paramétrique qui correspond ; toutefois, si le test d’hypothèse que vous utilisez n’est pas sensible à la normalité, vous pouvez l’utiliser même lorsque vos données ne sont pas normales.

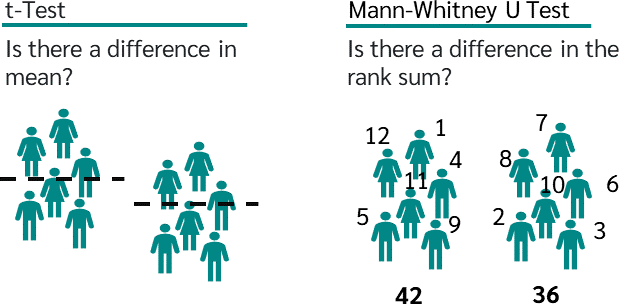
 QUELS SONT LES TESTS ROBUSTES À L'HYPOTHÈSE DE NORMALITÉ ?

Plusieurs tests sont «robustes» à l'hypothèse de normalité : les tests t de Student (1 échantillon, 2 échantillons et tests t appariés), l’analyse de la variance (ANOVA), la régression et les plans d’expériences (DOE).

**TEST U de Mann Whitney**

Le test U de Mann-Whitney peut être utilisé pour vérifier s'il existe une différence entre deux échantillons (groupes), et les données ne doivent pas nécessairement être distribuées normalement.

Pour déterminer s'il existe une différence entre deux échantillons, on utilise les sommes des rangs des deux échantillons plutôt que les moyennes comme dans le [test t pour les échantillons indépendants](https://datatab.fr/tutorial/unpaired-t-test).

****

**Interprétation des résultats du test U Mann-Whitney (non paramétrique du test t de Student)**

MannwhitneyuResult(statistic=5929.0, pvalue=9.080913416794144e-27) interprète moi les résultats du test de mann-whitney

ChatGPT

**ChatGPT**

Le test de Mann-Whitney est un test statistique non paramétrique utilisé pour déterminer s'il existe une différence significative entre les distributions de deux échantillons indépendants. La sortie du test de Mann-Whitney est une paire de valeurs : le statistique U et la valeur p.

* **Statistique U :** Dans ce cas, la valeur de la statistique U est 5929.0. La statistique U mesure la somme des rangs des observations de l'un des échantillons. Plus la valeur de U est grande, plus il est probable que les valeurs dans l'un des échantillons soient plus grandes que dans l'autre échantillon.
* **Valeur p :** La valeur p est 9.080913416794144e-27, ce qui est très proche de zéro. La valeur p représente la probabilité que les échantillons proviennent de la même distribution. Une valeur p faible (inférieure au seuil de signification généralement choisi, tel que 0.05) suggère qu'il y a des preuves suffisantes pour rejeter l'hypothèse nulle (l'hypothèse selon laquelle les distributions des échantillons sont égales) au profit de l'hypothèse alternative (l'hypothèse selon laquelle les distributions sont différentes).

Dans ce cas précis, avec une très petite valeur p (9.08e-27 << 0.05), cela suggère qu'il existe une différence statistiquement significative entre les deux groupes comparés. Cela signifie que les distributions des échantillons ne sont probablement pas les mêmes, bien que le test en lui-même ne dise pas dans quelle direction se situe cette différence.

***REGRESSION LINEAIRE***

**Théorie : la Régression Linéaire est basée sur les moindres carrés. Une relation linéaire implique une relation constante de ligne droite tel que y = x.**

**Pour minimiser l’erreur quadratiques : on utilise la méthode MCO.**

**Théorie des MCO de la régression linéaire :** y = mx + b (il n’y a de la place que pour une seule feature/caractéristique), les MCO permettront de résoudre les valeurs pour la pente m et pour l’interception b.

* Plus tard nous aurons besoin d’outils tels que la descente de gradient pour adapter cette méthode à de multiples features/caractéristiques

**Cas pratique :**

Features : x1, x2, x3, …, xn et y : que j’essaye de prédire

**But :** créer une relation linéaire entre les features X et le label y

**TEST DU CHI-2 (chi-carré)**

Le test du chi-2 (chi-carré) est un test statistique qui évalue si les différences entre les fréquences observées dans un tableau de contingence et les fréquences qui seraient attendues sous l'hypothèse nulle sont suffisamment importantes pour rejeter cette hypothèse.

Dans le contexte d'un tableau de contingence à deux dimensions (comme dans le cas d'un test du chi-2 pour l'indépendance entre deux variables catégorielles), le tableau des fréquences attendues représente les valeurs théoriques que l'on s'attendrait à observer dans chaque cellule du tableau si les variables étaient indépendantes.

Dans le contexte de votre test du chi-2, le tableau des fréquences attendues est le tableau qui aurait été observé si les variables étaient indépendantes. Chaque cellule de ce tableau représente la fréquence attendue pour une combinaison particulière des niveaux des variables catégorielles. Ces valeurs sont comparées aux fréquences observées pour évaluer si les différences sont significatives et remettent en question l'hypothèse d'indépendance entre les variables.

COURBE DE LORENZ

Cette visualisation représente ce qu'on appelle la "Courbe de Lorenz", un outil utilisé en économie pour illustrer la répartition de la richesse ou du revenu au sein d'une population. Voici une interprétation des résultats :

1. **Courbe de Lorenz** :
   * La courbe bleue représente la Courbe de Lorenz. Cette courbe mesure la distribution cumulative de la richesse dans une population. L'axe horizontal (proportion de la population) va de 0 à 1, indiquant la totalité de la population, tandis que l'axe vertical (proportion du chiffre d'affaires ou de la richesse) indique la part cumulée de la richesse détenue par cette proportion de la population.
2. **Égalité parfaite** :
   * La ligne orange représente l'égalité parfaite. Si la richesse était répartie de manière parfaitement égale, la courbe de Lorenz coïnciderait avec cette ligne. Dans ce cas, la distribution serait linéaire, montrant que chaque portion égale de la population détient la même part de richesse.
3. **Bissectrice** :
   * La ligne en pointillés grise est la bissectrice. Elle représente l'égalité absolue où chaque portion égale de la population détiendrait exactement la même part de richesse. Si la courbe de Lorenz se situe au-dessus de la bissectrice, cela indique une inégalité dans la répartition de la richesse.

En examinant cette courbe de Lorenz, vous pouvez interpréter le degré d'inégalité dans la répartition de la richesse. Plus la courbe de Lorenz s'éloigne de la ligne d'égalité parfaite, plus grande est l'inégalité de richesse dans la population étudiée.

Dans ce contexte, une courbe de Lorenz qui s'éloigne fortement de la ligne d'égalité parfaite indique une concentration plus élevée de la richesse entre les mains d'une fraction plus restreinte de la population, tandis qu'une courbe plus proche de la ligne d'égalité parfaite suggère une répartition plus équitable de la richesse.

**COEFFICIENT DE GINI**

Le coefficient de Gini est une mesure statistique utilisée pour évaluer le degré d'inégalité dans la distribution de la richesse ou du revenu au sein d'une population. Il varie de 0 à 1, où :

* 0 représente une égalité parfaite, signifiant que chaque individu détient la même part de richesse ou de revenu.
* 1 correspond à une inégalité totale, où une seule personne possède toute la richesse ou le revenu, et les autres n'ont rien.

Dans votre cas, le coefficient de Gini est de 0.4411. Cela indique qu'il y a une certaine inégalité dans la répartition de la richesse ou du chiffre d'affaires parmi la population étudiée. Plus précisément, environ 44.11 % de l'espace entre la courbe de Lorenz et la ligne d'égalité parfaite est observé, ce qui représente l'inégalité dans la distribution de la richesse.

En d'autres termes, cette valeur de coefficient de Gini suggère qu'il existe une inégalité modérée à substantielle dans la répartition de la richesse ou du chiffre d'affaires parmi les individus de la population étudiée, où une proportion significative de la richesse est détenue par une fraction relativement restreinte de la population.

**Moyenne mobile**

La moyenne mobile est une technique d'analyse de séries temporelles qui consiste à calculer la moyenne des valeurs d'une série sur une fenêtre glissante. Dans ce cas, le graphique montre le chiffre d'affaires mensuel ainsi que la moyenne mobile de ces données.

Voici une interprétation générale du calcul de la moyenne mobile :

1. **Calcul de la moyenne mobile** :
   * La moyenne mobile est une moyenne calculée sur une période spécifique, souvent utilisée pour lisser les variations à court terme et mettre en évidence les tendances à plus long terme dans les données de séries temporelles.
   * Dans ce cas, la moyenne mobile est calculée pour le chiffre d'affaires mensuel. Cela signifie que, pour chaque point de données, une moyenne est calculée sur une certaine fenêtre (par exemple, sur les derniers mois).
   * La colonne 'moyenne\_mobile' dans le DataFrame 'monthly\_revenue' semble contenir ces valeurs de moyenne mobile pour les chiffres d'affaires mensuels sur une période spécifiée.
2. **Tracé du graphique** :
   * Le graphique affiche le chiffre d'affaires mensuel en bleu, représentant les valeurs brutes du chiffre d'affaires pour chaque mois.
   * La ligne rouge représente la moyenne mobile calculée pour ces données de chiffre d'affaires mensuel. Cette ligne est moins susceptible de présenter des fluctuations brusques, car elle représente une moyenne sur une période donnée.
3. **Interprétation** :
   * En comparant la ligne bleue (chiffre d'affaires mensuel) à la ligne rouge (moyenne mobile), vous pouvez évaluer comment les valeurs fluctuent par rapport à la tendance moyenne sur cette période spécifique.
   * Si la ligne rouge est plus lisse que la ligne bleue, cela indique que les variations à court terme ont été atténuées, mettant davantage en évidence la tendance générale des chiffres d'affaires.

En résumé, la moyenne mobile est utilisée ici pour lisser les variations à court terme dans les données mensuelles de chiffre d'affaires, offrant ainsi une perspective plus claire sur la tendance sous-jacente des revenus au fil du temps.

**TEST ANOVA**

Le test ANOVA (Analyse de la variance) est utilisé pour évaluer s'il existe des différences significatives entre les moyennes de plusieurs groupes ou catégories. Dans votre cas, le test ANOVA a produit une statistique de test F de 41313.308 avec une valeur p (p-value) de 0.0, ce qui indique une différence significative entre les moyennes d'âge des différentes catégories.

Interprétation du test ANOVA :

* **Statistique de test F :** La statistique de test F mesure la variation entre les moyennes des différents groupes par rapport à la variation à l'intérieur des groupes. Une valeur élevée de la statistique F indique une différence significative entre au moins deux des moyennes des groupes.
* **Valeur p :** La valeur p est utilisée pour déterminer la signification statistique de la statistique de test F. Dans votre cas, une valeur p de 0.0 (probabilité très proche de zéro) indique une forte preuve contre l'hypothèse nulle, suggérant qu'il existe des différences significatives entre au moins deux des moyennes d'âge des catégories.
* **Conclusion :** Avec une valeur p inférieure à 0.05 (niveau de signification généralement utilisé), on rejette l'hypothèse nulle selon laquelle toutes les moyennes des groupes sont égales. Cela suggère donc qu'il y a des différences statistiquement significatives entre les moyennes d'âge des différentes catégories testées.

En résumé, le test ANOVA indique que les moyennes d'âge varient de manière significative entre les différentes catégories. Cela peut être utile pour déterminer quelles catégories spécifiques présentent des différences dans l'âge moyen des individus.