Les tests paramétriques et non paramétriques

I. Introduction

➢ *Qu'est-ce que la statistique ?*

La **statistique** est la discipline qui étudie des phénomènes à travers la collecte de *données*, leur traitement, leur analyse, l'interprétation des résultats et leur *présentation* afin de rendre ces données compréhensibles par tous. C'est à la fois une branche des mathématiques appliquées, une *méthode* et un ensemble de *techniques*.

La statistique est un domaine des mathématiques et de plus en plus, elle fait partie de ce que l'on appelle aujourd'hui la *science des données* (en anglais : *Data Science*).

Les statistiques sont un ensemble d'outils utilisés pour organiser et analyser des données.

➢ *Rôle des statistiques en écologie*

- Résumé l'information contenue dans un grand nombre d'observations
- Se convaincre et convaincre les autres

➢ *Qu'est-ce qu'un test statistique ?*

En statistiques, un **test**, ou **test d'hypothèse**, est une procédure de décision entre deux hypothèses.

Exemple : J'ai 114 livres dans ma bibliothèque. J'en extrais un échantillon de 12. Chaque livre doit avoir la même probabilité d'être choisi. Je veux tester l'hypothèse que la médiane du nombre de pages par volume est 220. Dans mon échantillon, j'observe les nombres de pages suivants :

126	142	156	228	245	246	370	419	433	454	478	503

Je leur associe un signe - si le nombre de pages est inférieur à 220 et un signe + sinon. Si la médiane est 220, il est également probable pour chaque livre sélectionné d'avoir plus ou moins de 220 pages.

En associant un + à une face et un - à un pile, nous pouvons faire une analogie avec un lancer de pièce. Nous verrons plus loin que le tirage "9 faces et 3 piles" a une probabilité assez forte pour qu'on ne puisse pas rejeter l'hypothèse selon laquelle le nombre de pages médian de mes livres est 220.

Exercice : Calculer la probabilité que parmi les 12 livres, 9 d'entre eux aient plus de plus de 220 pages, sous l'hypothèse que la médiane du nombre de pages des livres de ma bibliothèque est égale à 220.

r	0	1	2	3	4	5	6	7	8	9	10	11	12
P	0.000	0.003	0.016	0.054	0.121	0.193	0.226	0.193	0.121	0.054	0.016	0.003	0.000

Probabilités binomiales $P,\ \text{pour } r \text{ signes } +,\ n = 12,\ p = \frac{1}{2}$

➢ *Hypothèses de test*

En premier lieu, nous devons formuler les hypothèses. L'hypothèse que nous voulons vérifier sera appelée hypothèse nulle et on la notera $\boldsymbol{H_0}$*. Dans l'exemple concernant le nombre de pages des livres de ma bibliothèque, nous poserons alors*

$$H_0 : \theta = 220$$

Où $\boldsymbol{\theta}$ *représente ici la médiane du nombre de page. Nous rassemblerons d'autre part l'ensemble des hypothèses alternatives sous* $\boldsymbol{H_1}$ *:*

$$H_1 : \theta \neq 220$$

Et nous parlerons de tester $\boldsymbol{H_0}$ *contre les alternatives bilatérales* $\boldsymbol{H_1}$ *(sous* $\boldsymbol{H_1}$*,* $\boldsymbol{\theta}$ *peut être inférieur ou supérieur à 220).*

L'hypothèse nulle *notée* ***H_0*** *:* En statistiques et en économétrie, est une *hypothèse* postulant l'égalité entre des paramètres statistiques (généralement, la moyenne ou la variance) de *deux échantillons* dont elle fait l'hypothèse qu'ils sont pris sur des populations équivalentes. Elle est toujours testée contre une hypothèse alternative qui postule soit la différence des données (test bilatéral), soit une inégalité (*plus petit que* ou *plus grand que*) entre les données (test unilatéral).

Est celle que l'on considère vraie a priori. Le but du test est de décider si ce a priori est crédible.

L'hypothèse alternative *notée* ***H_1*** *:*

Est l'hypothèse complémentaire de **H_0**.

Exemple : Sous H0, le médicament n'a pas d'influence, sous H1 il en a une.

Attention

- *Les deux hypothèses ne sont pas symétriques. H1 est choisie uniquement par défaut si* ***H_0*** *n'est pas considérée comme crédible.*
- *Le choix de* ***H_0*** *et de* ***H_1*** *est en général impose par le test qu'on utilise et ne relève donc pas de l'utilisateur.*

Soit $\mu 1$ et $\mu 2$ les moyennes de tension des deux populations correspondant à la prise de médicament ou de placebo. Une manière de démontrer que le médicament modifie la tension est de montrer que μ_2 est différent de μ_1.

Les hypothèses deviennent alors {**H_0** : les moyennes des deux populations sont égales} et {**H_0** : les moyennes des deux populations sont différentes}. On l'écrit succinctement sous la forme :

$$H_0 : \quad \mu_1 = \mu_2$$
$$H_1 : \quad \mu_1 \neq \mu_2$$

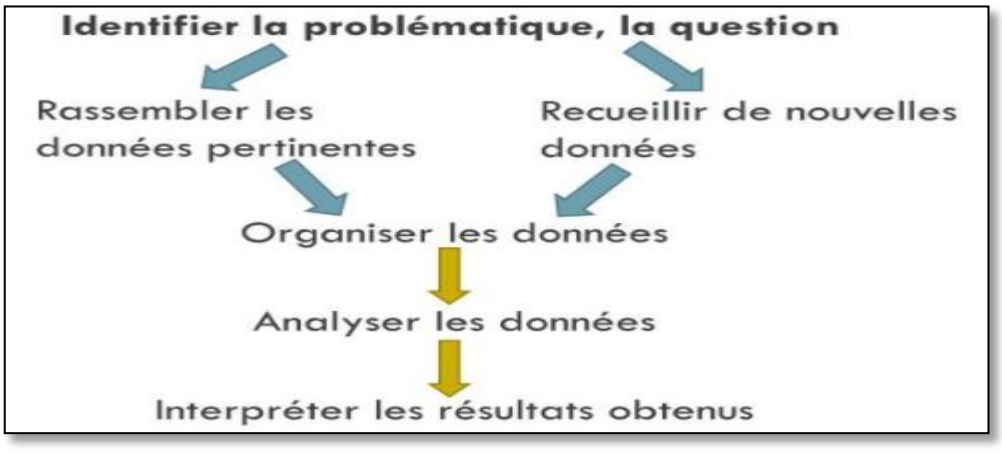
- ***La statistique de test***

La *statistique de test* ***S*** est une fonction qui résume l'information sur l'échantillon qu'on veut tester. On la choisit de façon à pouvoir calculer sa loi sous ***H_0***.

- **S** est une variable aléatoire, définie indépendamment des données observées. La valeur que prend cette variable aléatoire pour les données observées sera appelée ***statistique observée*** et notée ***S_{obs}*** dans la suite.
- Suivant le type de statistique choisi, le test sera ***paramétrique*** ou ***non-paramétrique***.

- ***Analyse de données***

1. Collecte des données ;
2. Synthèse et traitement ;
3. Présentation et communication de l'information.



➢ *Analyses statistiques : quels sont les grands types ?*

❖ *Statistique descriptive*

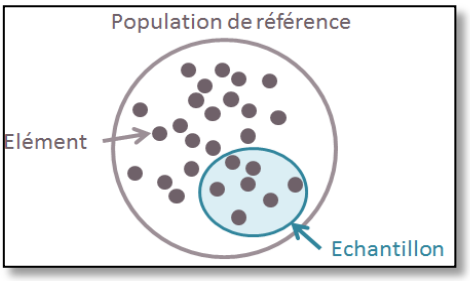
- Ensemble des outils et procédures qui permettent de réorganiser un ensemble de données, de les décrire et de les résumer.
- Il existe de nombreux moyens de représenter en tableau et graphiquement le données . Nous nous limiterons à décrire les distributions de fréquences et les histogrammes.

❖ *Statistique inférentielle*

- Statistique inférentielle consiste à extrapoler les résultats obtenus sur un échantillon afin de tirer des conclusions concernant la population.
- Il s'agit d'*inférer* certaines caractéristiques de la **population** à partir de ce que l'on sait sur un **échantillon.**

➢ *Population / Échantillon*

- Une **population** est un ensemble fini d'objets (les individus ou unités statistiques) sur lesquels une étude porte et dont les éléments répondent à une ou plusieurs caractéristiques communes.
- Un **échantillon** est un ensemble d'individus extraits d'une population étudiée de manière à ce qu'il soit représentatif de cette population, au moins pour l'objet de l'étude.



Exemple

➨ Supposez un écologiste qui voudrait mesurer la concentration des cyanobactéries dans les **lacs** de la région est du Québec.

↓ population

➨ Il/elle décide donc de mesurer la concentration de cyano dans **dix lacs.**

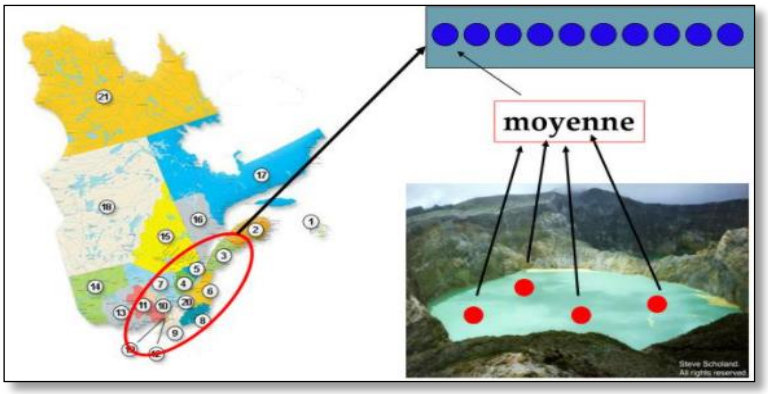
↓ échantillons

- Un **échantillon** (dix lacs) est tiré aléatoirement de la population.
- **Observation:** mesure une ou plusieurs propriétés (pH, cyanobactéries, taille etc...) d'une entité (ex.: lac).
- Le but est d'extrapoler les résultats à l'ensemble de la population.

Observations (4) et variables (2)

lac	pH	cyanobacterie (µg/L)
1	6.3	6.8
2	7.2	10.3
3	6.8	11.3
4	7.4	9.4

Une variable est une propriété d'une observation qui peut prendre différentes valeurs.



Suivant le type de statistique choisi, le test sera ***paramétrique*** ou ***non-paramétrique***.

II. Test paramétrique et non paramétrique

II.1. Test paramétrique

Un ***test paramétrique*** est un test pour lequel on fait une hypothèse paramétrique sur la distribution des données sous H_0 (distribution ***normale***, distribution de ***Poisson***...); Les hypothèses du test concernent alors les paramètres de cette distribution.

➢ En théorie des probabilités et en statistique, les ***lois normales*** (*lois gaussiennes, lois de Gauss*) sont parmi les lois de probabilité les plus utilisées pour modéliser des phénomènes naturels issus de plusieurs événements aléatoires.

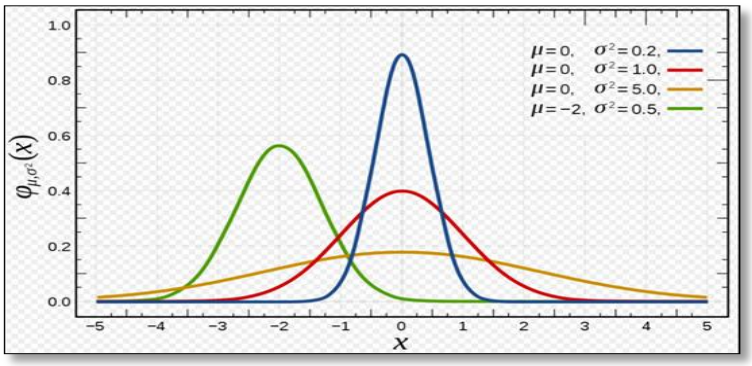
Plus formellement, une loi normale est une loi de probabilité absolument continue qui dépend de deux paramètres : son espérance, un nombre réel noté μ, et son écart type, un nombre réel positif noté σ.

La ***densité de probabilité*** de la loi normale ***d'espérance μ***, et d'***écart type σ*** est donnée par :

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

- En théorie des probabilités, *l'**espérance mathématique*** d'une variable aléatoire réelle est, intuitivement, la valeur que l'on s'attend à trouver, en moyenne, si l'on répète un grand nombre de fois la même expérience aléatoire. Elle se note E (x) et se lit « espérance de x ».
- En mathématiques, *l'**écart type*** est une mesure de la dispersion des valeurs d'un échantillon statistique ou d'une distribution de probabilité. Il est défini comme la racine carrée de la variance ou, de manière équivalente, comme la moyenne quadratique des écarts par rapport à la moyenne.

Densité de probabilité
*La courbe rouge représente la fonction **φ**, densité de probabilité de la loi normale centrée réduite.*
[La fonction de densité de la loi normale centrée réduite (dite courbe de Gauss ou courbe en cloche)].



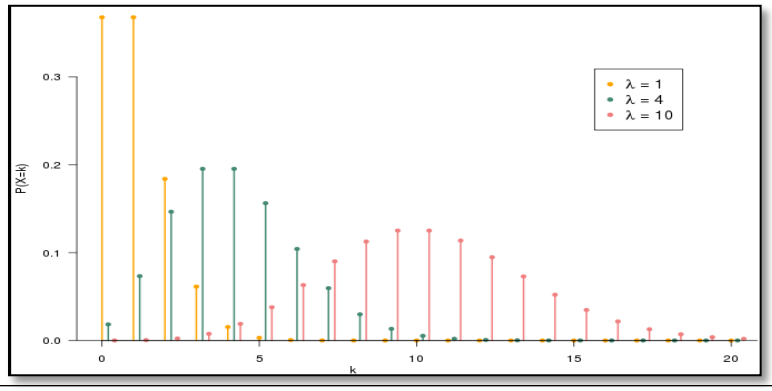
➢ En théorie des probabilités et en statistiques, ***la loi de Poisson*** est une loi de probabilité discrète qui décrit le comportement du nombre d'événements se produisant dans un intervalle de temps fixé, si ces événements se produisent avec une fréquence moyenne ou espérance connue, et indépendamment du temps écoulé depuis l'événement précédent.

Par **exemple**, si un certain type d'événements se produit en moyenne 4 fois par minute, pour étudier le nombre d'événements se produisant dans un laps de temps de 10 minutes, on choisit comme modèle une **loi de Poisson** de paramètre λ = 10×4 = 40.

La variable aléatoire ***X*** prend des valeurs positives entières ***k*** (par exemple des unités de temps 1, 2, 3…).

$$\boldsymbol{P(X = k) = \frac{e^{-\lambda}\lambda^{k}}{k!}}$$

Fonction de masse
Les fonctions de masse ne sont définies que pour les entiers k.



II.1.1. T-test (test-t de Student)

Probablement l'une des questions les plus populaires dans le domaine de la recherche est de savoir si **deux groupes d'échantillons indépendants** diffèrent l'un de l'autre. Le **test de Student** est l'un des tests statistiques le plus utilisé pour comparer les moyennes de deux **groupes indépendants** ou **appariés**.

Le **test-t de Student** est un **test statistique** permettant de **comparer les moyennes** de deux groupes d'échantillons. Il s'agit donc de savoir si les moyennes des deux groupes sont significativement différentes au point de vue **statistique**.

Il existe plusieurs variants du **test-t de Student**:

- Le test-t de Student pour **échantillon unique**
- Le test-t de Student comparant deux groupes **d'échantillons indépendants** (on parle de **test de Student non apparié**)
- Le test-t de Student comparant deux groupes **d'échantillons dépendants** (on parle de **test de Student apparié**).

L'objectif ici et de décrire les formules pour les différents types de **test de Student**. Le **test de Student** est dit **paramétrique** car, comme nous allons le voir, la formule dépend de la **moyenne** et de **l'écart-type** des observations à comparer.

❖ Test de Student pour échantillon unique

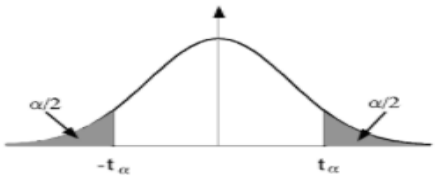
Il s'agit de comparer une **moyenne observée** à une **moyenne théorique** (μ). Soit X une série de valeurs de taille n, de **moyenne** m et d'**écart-type** S. La comparaison de la **moyenne observée** (m) à une **valeur théorique** μ est donnée par la formule :

$$t = \frac{m - \mu}{s/\sqrt{n}}$$

Pour savoir si la différence est significative, il faut tout d'abord lire dans la **table t**, la valeur critique correspondant au **risque alpha** = 5% pour un degré de liberté :

$$d.d.l = n - 1$$

- La **table de Student** ou **table t** donne la probabilité **alpha** (**α**) pour que **t** égale ou dépasse, en valeur absolue, une valeur donnée, en fonction du nombre de degrés de liberté **(d.d.l.).**
- En statistiques le **degré de liberté** (**d.d.l**) désigne le nombre de variables aléatoires qui ne peuvent être déterminées ou fixées par une équation (notamment les équations des tests statistiques).



Exemple : avec d.d.l. =10, pour t=2.228 la probabilité est alpha=0.05

d.d.l/α	0.9	0.5	0.3	0.2	0.1	0.05	0.02	0.01	0.001
1	0.158	1	2	3.078	6.314	12.706	31.821	64	637
2	0.142	0.816	1.386	1.886	2.92	4.303	6.965	10	31.598
3	0.137	0.765	1.25	1.638	2.353	3.182	4.541	5.841	12.929
⋮									
40	0.126	0.681	1.05	1.303	1.684	2.021	2.423	2.704	3.551
80	0.126	0.679	1.046	1.296	1.671	2	2.39	2.66	3.46
120	0.126	0.677	1.041	1.289	1.658	1.98	2.358	2.617	3.373
Infini	0.126	0.674	1.036	1.282	1.645	1.96	2.326	2.576	3.291

! *Le test n'est applicable que si seulement si la série de valeurs X suit une loi normale.*

❖ *Test t de Student pour échantillons indépendants*

Il s'agit de **comparer deux moyennes observées**. Lorsque les deux groupes d'échantillons (A et B) à comparer n'ont aucun lien, on utilise le **test t de Student indépendant** (ou **non apparié**).

C'est quoi le test de Student non-apparié ?

A titre d'exemple, nous avons un groupe de 100 individus (50 femmes et 50 hommes) pris au hasard au sein de la population. *On se pose la question à savoir si le poids moyen des femmes est significativement différent de celui des hommes?*

Dans cet exemple on parle de **test de Student non apparié** car les deux groupes à comparer n'ont aucun lien. Il s'agit donc de calculer le poids moyen des femmes et de celui des hommes et d'évaluer si la **différence est significative** au point de vue **statistique**.

Formule

- Soit A et B deux groupes différents à comparer.
- Soit m_A et m_B la moyenne du groupe A et celui du groupe B, respectivement.
- Soit n_A et n_B la taille du groupe A et celle du groupe B, respectivement.

La **valeur t de Student** est donnée par la formule suivante:

$$t = \frac{m_A - m_B}{\sqrt{\frac{S^2}{n_A} + \frac{S^2}{n_B}}}$$

S^2 est la **variance** commune aux deux groupes. Elle est calculée par la formule suivante :

$$S^2 = \frac{\sum (x - m_A)^2 + \sum (x - m_B)^2}{n_A + n_B - 2}$$

Pour savoir si la différence est significative, il faut tout d'abord lire dans la table t, la valeur critique correspondant au **risque alpha** = 5% pour un degré de liberté :

$$d.d.l = n_A + n_B - 2$$

- **Le risque de première espèce ou risque α**

C'est le risque de rejeter l'hypothèse nulle (H_0) alors que celle-ci est vraie. Ce risque est parfaitement connu : c'est la probabilité utilisée lors de la réalisation du test, c'est à dire la probabilité pour que la valeur de la variable aléatoire Z (distribuée selon la loi de distribution choisie pour le test) soit extérieure à l'intervalle [z_1, z_2] ([-2,093 ;+2.093] dans notre exemple pour une loi de Student, 19 ddl et un risque α égal à 5 %). La valeur 5% est classiquement utilisée avec dans certains cas l'utilisation d'un risque 1%.
Au risque 5 % on estime donc que la probabilité pour que la différence observée soit due aux fluctuations d'échantillonnage est suffisamment faible pour accepter H_0.

- **p-valeur**, est la probabilité pour un modèle statistique donné sous l'hypothèse nulle d'obtenir la même valeur ou une valeur encore plus extrême que celle observée.

! *Si la valeur absolue de t (|t|) est supérieure à la valeur critique, alors la différence est significative. Dans le cas contraire, elle, ne l'est pas. Le degré de significativité ou p-value correspond au risque indiqué par la table de Student pour la valeur |t|*

!! *- Le test est utilisable, si seulement si, A et B suivent des **lois normales** de mêmes variances.*
*- Lorsque les variances des deux groupes à comparer sont différentes, le **test t de Welch est préconisé**.*

❖ Test-t de Student pour séries appariés

C'est quoi le test de Student apparié ?

Le **test de Student apparié** permet de comparer la moyenne de deux séries de valeurs ayant un lien.
Par exemple, 20 souris ont reçu un traitement X pendant 3 mois. On se pose la question à savoir si le traitement X a un impact sur le poids des souris au bout des 3 mois. Le poids des 20 souris a donc été mesuré avant et après traitement.
Ce qui nous donne 20 séries de valeurs avant traitement et 20 autres séries de valeurs après traitement provenant de la mesure du poids des mêmes souris.

Il s'agit bien dans cet exemple, d'un **test de Student apparié** car les deux séries de valeurs ont un lien (les souris). Pour chaque souris, on a deux mesures (l'une avant et l'autre après traitement).

Formule

Pour **comparer les moyennes de deux séries appariées**, on calcule tout d'abord la différence des deux mesures pour chaque paire.

Soit d la série des valeurs correspondant aux différences des mesures entre les paires de valeurs. La moyenne de la différence d est comparée à la valeur 0. S'il y a une différence significative entre les deux **séries appariées**, la moyenne de d devrait être très éloignée de la valeur 0.

La **valeur t de Student** est donnée par la formule :

$$t = \frac{m}{s/\sqrt{n}}$$

m et **s** représentent la **moyenne** et l'**écart-type** de la différence d. **n** est la taille de la série d.

Pour savoir si la différence est significative, il faut tout d'abord lire dans la **table t**, la valeur critique correspondant au **risque alpha** = 5% pour un degré de liberté :

$$d.d.l = n - 1$$

! *Si la valeur absolue de t (|t|) est supérieure à la valeur critique, alors la différence est significative. Dans le cas contraire, elle, ne l'est pas. Le degré de significativité (ou p-value) correspond au risque indiqué par la table de Student pour la valeur |t|.*

!! *Le test est utilisable, si seulement si, la différence de suit une loi normale.*

❖ *Test de Student en ligne*

Vous n'avez plus besoin de logiciel SPSS (Statistical Package for the Social Sciences / Ensemble des programmes statistiques pour les sciences sociales) ou d'Excel pour faire le **test t de Student**.

Un logiciel web est disponible ***(http://www.sthda.com/french/rsthda/rsthda.php)*** pour faire le test de Student en ligne sans aucune installation. Vous devez juste être membre du site et être connecté.

En fonction du type de **test de Student** que vous souhaitez faire, cliquez sur les liens suivant :

- Test-t de Student pour échantillon unique :
  http://www.sthda.com/french/rsthda/one-sample-t-test.php
- Test-t de Student pour échantillon indépendant :
  http://www.sthda.com/french/rsthda/unpaired-t-test.php
- Test-t de Student pour échantillon apparié :
  http://www.sthda.com/french/rsthda/paired-t-test.php

Principe du test

Les étapes d'un test sont toujours réalisées dans l'ordre suivant :

1) Choix du risque α
2) Choix du type de test et de sa latéralité si besoin
3) Calcul de la statistique de test
4) Calcul de la p-valeur
5) Conclusion

En pratique, l'utilisation d'un logiciel type R (ver. 3.1.0) permet de ne pas se soucier des parties 3) et 4). Par contre, les choix liés aux étapes 1) et 2) ainsi que l'interprétation finale ne peuvent être faits par le logiciel.

II.1.2. ANOVA simple & multiple

- ***Concepts de base de l'ANOVA***

L'analyse de variance (expression souvent abrégée par le terme anglais ANOVA : **An**alysis **O**f **Va**riance) est une technique statistique. Son but est en fait de procéder à une sorte de généralisation de la comparaison des moyennes ou de la comparaison des pourcentages lorsqu'il y a plus de deux valeurs à comparer.

Soit un ensemble d'individus (par exemple des élèves, des établissements, etc…) caractérisables par une ou plusieurs variables qualitatives (par exemple l'origine sociale des élèves, le statut public ou privé des établissements…) et quantitatives (par exemple les notes des élèves, les effectifs des établissements…). L'analyse de variance vise à tester si, pour une variable qualitative donnée, le fait que des individus appartiennent à des catégories différentes de cette variable exerce ou non un effet statistiquement significatif sur les caractéristiques de ces individus en termes de telle ou telle variable quantitative.

On appelle ***variable indépendante*** ou "***facteurs de variabilité***" (ou simplement "***facteurs***") ***les caractéristiques qualitatives*** (et leurs catégories) dont on étudie l'impact potentiel sur ***les caractéristiques quantitatives***. Ces dernières sont ***les variables dépendantes*** (ou « ***variables-réponses*** »).

Le ***but de l'analyse de la variance*** est de ***tester la présence de différences significatives ou non entre des moyennes***. L'analyse de la variance est la méthode employée pour tester plusieurs moyennes.

Dans une anova, on considère une unique variable dépendante. L'anova elle-même recouvre deux grands cas de figure, ***l'anova à un seul facteur*** et ***l'anova multifactorielle***, cette dernière comportant elle-même plusieurs possibilités : l'anova à deux facteurs, l'anova à trois facteurs, etc.

a. ANOVA À UN FACTEUR

On peut classiquement introduire l'analyse de la variance ou ***l'ANOVA***(***AN***alysis ***O***f ***VA***riance) de deux manières.

- On cherche à expliquer une variable ***quantitative Y*** au moyen d'un variable explicative ***qualitative X.***

Généralement, la variable explicative est appelée ***facteur*** (***explicatif***). Les ***modalités*** sont appelées ***niveaux de facteur***

- On désire comparer différentes populations ou différentes conditions expérimentales. La question que l'on se pose est : " Y-a-t-il des différences en moyenne entre les divers groupes ?"

Exemple1 : Un agronome veut étudier l'effet de 3 types d'engrais sur le rendement à l'hectare de parcelles de blé. Ici,

Y = Rendement à l'hectare (en tonnes), variable quantitative,
X = type d'engrais (A, B ou C), variable qualitative ou facteur.

Quel modèle linéaire pour étudier cette problématique ?

Reprenons l'exemple des engrais. Supposons que ***ni*** observations du rendement à l'hectare ont été obtenues pour chaque type d'engrais ***i***. Un modèle peut s'écrire sous la forme :

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

Avec i = A, B, C et j = 1,…… ni.

- Y_{ij} est le rendement observé de la parcelle j traitée avec l'engrais i ;
- μ_i représente le rendement moyen d'une parcelle traitée avec l'engrais i
- ε_{ij} est un terme d'erreur aléatoire.

Ce modèle peut aussi s'écrire sous la forme alternative suivante :

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

où μ est le rendement moyen global (quel que soit l'engrais) et α_i est l'effet différentiel du niveau i du facteur "engrais".

Une *hypothèse intéressante* à tester ici est :

$$H_0 : \mu_A = \mu_B = \mu_C$$

Ou de manière équivalente :

$$H_0 : \alpha_A = \alpha_B = \alpha_C = 0.$$

Le but sera donc :

- tester si deux niveaux différents du facteur entraînent une différence significative dans la variable à expliquer Y.
- ou de manière équivalente tester l'effet du facteur X ("engrais") sur la variable à expliquer Y ("rendement"),

Représentation des données

→ On suppose que l'on dispose de ***k*** échantillons de tailles respectives ***n_1,……, n_k***, correspondant aux ***k niveaux d'un facteur***. La taille de la population est donc :

$$n = \sum_{i=1}^{k} n_i .$$

→ Les variables aléatoires de notre n échantillon (***Y_{ij}***) sont alors *indicées par 2 dimensions* ($1 \leq i \leq k$ est le ***niveau*** d'appartenance de ***Yij*** et $1 \leq j \leq n_i$ est le ***numéro*** d'apparition de ***Y_{ij}*** dans le niveau ***i***). Il s'ensuit que ***n_i*** est " ***l'effectif*** " du niveau ***i***.

→ Nous pouvons ainsi représenter nos données sous forme de tableau :

Niveaux	Effectifs	Variables à expliquer
1	n_1	$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$
2	n_2	$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$
⋮	⋮	⋮
k	n_k	$Y_{k1}, Y_{k2}, \ldots, Y_{kn_k}$

→ ***Par exemple,*** le tableau ci-dessous donne le rendement en quintaux par hectare d'une variété de blé cultivée avec les engrais A, B et C.

Niveaux	Effectifs	Rendements
A	4	48, 49, 50, 49
B	4	47, 49, 48, 48
C	4	49, 51, 50, 50

L'agriculteur se demande :

- si le type d'engrais (A, B ou C) a un effet sur le rendement moyen de la variété de blé qu'il cultive ?
- quel type d'engrais est lui permet d'obtenir un meilleur rendement à l'hectare ?

=> La variable explicative qualitative (le facteur) définit des ***groupes*** (classes) d'appartenances des variables à expliquer ***Yij***.

Notations

→ Au niveau de chaque groupe ***i*** (pour ***i* = *1,….., k***),

$$Y_{i.} = \sum_{j=1}^{n_i} Y_{ij} \quad \text{et} \quad \bar{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

représentent ***la somme*** ainsi que la ***moyenne empirique*** du niveau ***i***.

→ Au niveau de l'ensemble des observations,

$$Y_{..} = \sum_{i=1}^{k} Y_{i.} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij} \quad \text{et} \quad \bar{Y}_{..} = \frac{1}{n}\sum_{i=1}^{k} Y_{i.} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$$

représentent ***la somme*** sur tous les niveaux ainsi que la ***moyenne empirique*** de l'échantillon global.

Modélisation

✓ La ***première modélisation*** est donnée par :

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad \text{pour } i = 1, \ldots, k \quad \text{et} \quad j = 1, \ldots, n_i \quad \text{, où :}$$

→ les erreurs ε_{ij} sont des variables aléatoires indépendantes, gaussiennes centrées, homoscédastiques de variance $\sigma^2 > 0$,

→ les observations Y_{ij} sont des variables aléatoires centrées autour d'une moyenne μ_i propre au niveau i avec :

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2) \quad \text{et que} \quad \mathrm{Cov}(Y_{ij}, Y_{k\ell}) = \begin{cases} \sigma^2 & \text{si } i = k \text{ et } j = \ell \\ 0 & \text{sinon.} \end{cases}$$

⇒Les variables aléatoires Y_{ij} sont donc indépendantes mais non identiquement distribuées (puisque leur espérance dépend de leur niveau d'appartenance).

→la moyenne théorique μ_i de la variable à expliquer Y dans le groupe de niveau i est appelée l'effet du niveau i du facteur X.

✓ La ***seconde modélisation*** est alors donnée par :

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \text{pour } i = 1, \ldots, k \quad \text{et} \quad j = 1, \ldots, n_i$$

où les erreurs (ε_{ij}) forment toujours un bruit blanc gaussien de variance $\sigma^2 > 0$ et donc

$$Y_{ij} \sim \mathcal{N}(\mu + \alpha_i, \sigma^2) \quad \text{avec} \quad \mathrm{Cov}(Y_{ij}, Y_{k\ell}) = \begin{cases} \sigma^2 & \text{si } i = k \text{ et } j = \ell \\ 0 & \text{sinon.} \end{cases}$$

→ L'interprétation de l'effet global μ et de l'effet marginal α_i dépendent du choix de la décomposition de $\mu_i = \mu + \alpha_i$,

b - ANOVA À DEUX FACTEURS

Également appelé *two-way ANOVA*, l'analyse de la variance à deux facteurs s'applique lorsque l'on souhaite prendre en compte deux facteurs de variabilité.

Par exemple, si on utilise deux engrais simultanément, on espère que l'action des engrais se complète et que les plantes se développent mieux du fait de cette concomitance.

On peut ainsi mesurer l'influence de divers dosages de chacun des engrais.

L'interaction peut être bénéfique (synergie) ou néfaste (antagonisme).

p valeurs pour le facteur ***F1***, ***q*** pour le facteur ***F2***, r observations pour chaque couple de facteur.

Le modèle est :

$$Y_{ijk} = m_{ij} + \gamma_{ij} + \epsilon_{ijk}$$

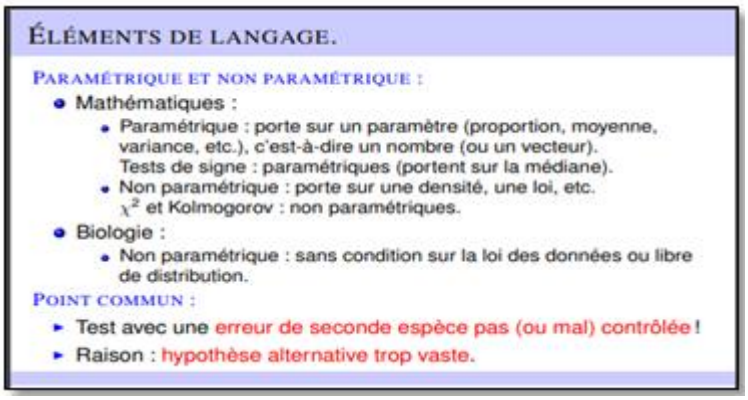
Que l'on écrite :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

Les ϵ_{ijk} sont des v.a. Normales, centrées et de même écart-type σ.

II.2. Test non paramétrique

Un ***test non paramétrique*** est un test ne nécessitant pas d'hypothèse sur la distribution des données. Les données sont alors remplacées par des statistiques ne dépendant pas des moyennes / variances des données initiales (tableau de contingence, statistique d'ordre comme les rangs...).



Test paramétrique ou non paramétrique ?

- *Test paramétrique* : les hypothèses nulle et alternative du test portent sur un paramètre statistique (moyenne ou variance par exemple). Ces tests nécessitent généralement des conditions de validité (distribution normale des données par exemple).
- *Test non paramétrique* : un test non paramétrique porte globalement sur la répartition des données sans hypothèse sur leur distribution (distribution free).

Données indépendantes ou appariées ?

- Données indépendantes : les observations sont indépendantes à l'intérieur de chaque échantillon et d'un échantillon à l'autre.

Ex: résultats scolaires filles et garçons, dosage d'un produit chez 2 groupes de patients ayant reçu une molécule ou un placebo...

- Données appariées : les mêmes individus sont soumis à 2 mesures successives d'une même variable.

Ex: notes de copies soumises à une double correction, dosage d'un produit avant et après un traitement chez les mêmes individus...

II.2. 1. Tests non paramétriques (Mann-Whitney, Wilcoxon, Kruskall-Wallis) de comparaison des caractères quantitatifs.

Lorsque les échantillons peuvent être considérés indépendants, on applique le test de Mann et Whitney pour 2 échantillons, celui de Kruskal et Wallis pour un nombre quelconque d'échantillons. Lorsque on a aaire à deux échantillons appariés (c'est-à-dire non indépendants), on applique le test de Wilcoxon.

Tous ces tests sont dits non paramétriques car ils ne nécessitent pas d'estimation de la moyenne et de la variance :

II.2. 1.1. Test de Mann et Whitney

On dispose des mesures des valeurs de X dans deux échantillons indépendants E_1 et E_2, de tailles respectives n_1 et n_2. On souhaite comparer les deux moyennes expérimentales, c'est-à-dire tester l'hypothèse nulle (H_0) : « $\mu_1 = \mu_2$ ».

On commence par trier les valeurs obtenues dans la réunion des deux échantillons par ***ordre croissant***. Pour chaque valeur x_i issue de E_1, on compte le nombre de valeurs issues de E_2 situées après lui dans la liste ordonnée (celles qui sont égales à x_i ne comptent que pour $\frac{1}{2}$). On note u_1 la somme des nombres ainsi associés aux déférentes valeurs issues de E_1. On fait de même en échangeant les rôles des deux échantillons, ce qui donne la somme u_2. Soit u la plus petite des deux sommes obtenue

$$u = min\{u_1 ; u_2\}$$

On note u la variable aléatoire associée.

- Pour n_1 et n_2 quelconques, on lit dans les tables du ***test de Mann et Whitney*** le nombre m_α tel que, sous (H_0), $P(u \leq m_\alpha) = \alpha$. On rejette ($H_0$) au risque d'erreur α si $u \leq m_\alpha$.

Autrement on accepte (H_0).

- Si n_1 et n_2 sont assez grands (≥ 20 en général), sous (H_0), u suit approximativement la loi normale $N(\mu; \sigma)$ avec :

$$\mu = \frac{n_1 n_2}{2} \quad \text{et} \quad \sigma = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \ .$$

On lit u_α dans la table de l'écart réduit de la loi normale tel que **P** ($|N| \geq u_\alpha$) = **α**, on calcule :

$$\varepsilon = \frac{u - \mu}{\sigma}$$

et on rejette (H_0) au risque d'erreur α si $\varepsilon \notin]-u_\alpha; u_\alpha[$. Autrement on accepte ($H_0$).

II.2. 1.2. Test de Kruskal et Wallis

On dispose des mesures des valeurs de X dans k échantillons indépendants E_1;..... ; E_k, de tailles respectives n_1;....; n_k. On souhaite comparer les k moyennes expérimentales, c'est-à-dire tester l'hypothèse nulle (H_0) : « $\mu_1 = \cdots = \mu_k$ ».

Pour chaque échantillon Ei, on calcule la somme ri des rangs des valeurs qui en sont issues. On pose :

$$h = \frac{12}{n(n+1)} \left(\sum_{i=1}^{k} \frac{r_i^2}{n_i} \right) - 3(n+1) \ ,$$

où $n = \sum_{i=1}^{k} n_i$ est l'effectif global. On note H la variable aléatoire associée.

- Si $n_1, \ldots, n_k$ sont assez grands (> 5 en général), sous (H_0), H suit approximativement la loi du χ^2 à $k-1$ degrés de liberté. On lit χ^2_α dans la table du χ^2 tel que $P(H \geq \chi^2_\alpha) = \alpha$ et on rejette (H_0) au risque d'erreur α si $h \geq \chi^2_\alpha$. Autrement on accepte (H_0).
- Dans le cas où on dispose de 3 échantillons de petite taille (≤ 5), on lit dans les tables du test de Kruskal et Wallis ci-dessous le nombre h_α tel que, sous (H_0), $P(H \geq h_\alpha) = \alpha$. On rejette (H_0) au risque d'erreur α si $h \geq h_\alpha$. Autrement on accepte (H_0).

II.2. 1.3. Test de Wilcoxon

On dispose de deux échantillons appariés E_1 et E_2, c'est-à-dire que chaque valeur de E_1 est associée à une valeur de E_2. On teste l'hypothèse nulle (H_0) : « $\mu_1 = \mu_2$ ».

On calcule les déférences entre les valeurs appariées, puis on les classe par ordre croissant des valeurs absolues, en omettant les déférences nulles. On affecte à chaque déférence non nulle son rang dans le classement (ou la moyenne de ses rangs en cas d'ex-æquo). On note w_+ la somme des rangs des différences strictement positives, w_- la somme des rangs des déférences strictement négatives ; on vérifie que :

$$w_+ + w_- = \frac{N(N+1)}{2} \ ,$$

Où N désigne les nombres de déférences non nulles. Enfin, on note w le plus petit des deux nombres : $w = \min\{w_+, w_-\}$ Soit W la variable aléatoire associée à w.

II.2.2. Tests non paramétriques de comparaison des caractères qualitatifs (Chi 2 simple et sur tableau de contingence).

II.2.2.1.Test du chi-deux

Les ***tests du chi-deux*** (ou ***khi-deux***, ou ***khi carré***, ou χ^2, ou ***Test du χ² de Pearson***) sont des tests d'hypothèses statistiques non-paramétriques. Ils tirent leur nom du fait que l'on lit l'écart critique dans la table de la loi du chi-deux. Ils sont essentiellement au nombre de trois :

1. Le test du chi-deux d'**ajustement**, ou d'**adéquation**, qui compare globalement la distribution observée dans un échantillon statistique à une distribution théorique.

2. le test du chi-deux d'**indépendance**, qui teste si deux caractères d'une population sont indépendants.
3. le test d'**homogénéité**, qui teste si des échantillons sont issus d'une même population.

Test d'ajustement du chi-deux

Le but de ce test est de comparer une distribution théorique d'un caractère à une distribution observée. Pour cela, le caractère doit prendre un nombre fini de valeurs, ou bien ces valeurs doivent être rangées en un nombre fini de classes.

- Données :
  - un caractère A dont les valeurs possibles sont réparties en k classes $A_1,\ldots,A_k$. La probabilité théorique dans chacune des classes est notée $p_1,\ldots,p_k$.
  - n observations, qui donnent un effectif n_1 pour la classe $A_1,\ldots,n_k$ pour la classe A_k. Bien sûr, on doit avoir $n_1+\ldots+n_k=n$.
- Hypothèse testée : "La distribution observée est conforme à la distribution théorique" avec un risque d'erreur a.
- Déroulement du test :
  1. On calcule les effectifs théoriques np_j.
  2. On calcule la valeur observée de la variable de test :

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j}$$

  3. On cherche la valeur critique χ^2_a dans la *table de la loi du chi-2* à k-1 degrés de liberté.
  4. Si $\chi^2 < \chi^2_a$, on accepte l'hypothèse, sinon on la rejette.
  5. Vérification a posteriori des conditions d'application : il faut $npj \geq 5$ pour tout j.

Test d'indépendance du chi-deux

Le test d'indépendance du chi-deux permet de contrôler l'indépendance de deux caractères dans une population donnée.

II.2.2.2.Tableau de contingence

Un tableau de contingence permet de représenter la répartition d'effectifs d'un échantillon en fonction de la valeur d'une observation, Cette valeur peut être discrète ou continue, dans ce cas elle sera modélisée en classes.

Exemple : 137 patients atteints de cirrhose sont divisés en 2 groupes. Le premier groupe C_1 reçoit un m´médicament, le second C_2 un placebo. Le stade d´évolution de la maladie constitut la variable observée et est définie en 3 modalités pour 3 stades d´évolution : $d_1 = 1$ $d_2 = 2$ $d_3 = 3$.

	Stade			
	1	2	3	Total
placebo	13	29	26	68
traitement	16	37	16	69
Total	29	66	42	137

	Stade			
	1	2	3	Total
placebo	0.191	0.426	0.382	1
traitement	0.232	0.536	0.232	1
Total	0.212	0.482	0.306	1

On peut voir que la proportion de patients en stade 3 pour les patients sous traitement est moins forte que celle des patients sous placebo.
Question : cette différence est-elle significative ?

III. Quelle est la différence entre un test paramétrique et un test non-paramétrique ?

1. *Les tests paramétriques* se basent sur des distributions statistiques supposées dans les données. Par conséquent, certaines conditions de validité doivent être vérifiées pour que le résultat d'un test paramétrique soit fiable. Par exemple, le test t de Student pour échantillons indépendants n'est fiable que si les données associées à chaque échantillon suivent une distribution normale et si les variances des échantillons sont homogènes.

Les tests non-paramétriques ne se basent pas sur des distributions statistiques. Ils peuvent donc être utilisés même si les conditions de validité des tests paramétriques ne sont pas vérifiées.

Les tests paramétriques ont souvent des tests non-paramétriques équivalents. Vous trouverez une série de tests paramétriques courants ainsi que leurs équivalents non-paramétriques, lorsqu'ils existent, dans cette grille.

2. *L'avantage d'utilisation*

Les tests non-paramétriques sont plus robustes que les tests paramétriques. En d'autres termes, peuvent être utilisés dans un plus grand nombre de situations.

Les tests paramétriques sont, eux, plus puissants en général que leurs équivalents non-paramétriques. En d'autres termes, un test paramétrique sera plus apte à aboutir à un rejet de H_0, si ce rejet est justifié. La plupart du temps, la ***p***-value calculée par un test paramétrique sera inférieure à la ***p***-value calculée par un équivalent non-paramétrique exécuté sur les mêmes données.

RÉFRENCES

- Analyse de la variance. - Chapitre VI - L3 MIASHS - Université de Bordeaux.p.37
- Compléments de Mathématiques. Tests non paramétriques. Licence de Biologie, 3e semestre. Faculté des Sciences et Techniques de Limoges.2007-2008. S. Vinatier. p.3
- Statistiques non-paramétriques.Ch.1. Tests non-paramétriques 2017-18.M2 CEE. Pr. Philippe Polomé, Université Lumière Lyon 2.p.92.
- Statistiques. Rappels de cours et travaux dirigés. Jean-Marc Labatte. Master 1.Biologie et technologie du végétal .Année 2010-2011. P.87
- Tests du chi-deux. Charles-Jean de La Vallée Poussin. *Bibm@th.net*