Sommaire

Principe d’un test statistique

Comprendre le principe général d’un test d’hypothèse

- Comprendre les règles de bon usage des tests statistiques
- Connaître les notions de
  - Risque de première espèce
  - Risque de deuxième espèce
    - Puissance d’un test
    - Calcul du nombre de sujets

Estimer les paramètres d’une population inconnue

- De tendance centrale
- De dispersion

- Comparer des paramètres observés sur plusieurs échantillons (populations)
  - Échantillon vs population
  - Plusieurs échantillons
- Etablir des modèles prédictifs

Deux possibilités pour décrire / comparer des populations

Tester des populations entières mais c'est exceptionnel

– Tester des échantillons et extrapoler à leurs populations source

une statistique = paramètres de l'échantillon (moyenne, écart-type, …)

• inférence statistique = porter une conclusion sur l'ensemble de la population source

Les tests statistiques sur échantillons n'ont d'intérêt que rapportés à leur population source

• Les tests statistiques paramétriques sont liés à des lois de probabilités

• associer une probabilité de survenue à tout événement

:

Les tests statistiques en pratique

Quelques questions:

dans une usine de produits chimiques, le Volume Globulaire Moyen de 30 ouvriers a été testé (92,5µm^3) et comparé à celui de 30 employés de bureau (94,7µm^3)

Avec un traitement A on observe 77% de guérison. Avec un traitement B 68 %.

Ces différences sont-elles réelles ou dues au hasard ?

Les tests statistiques substituent à une solution empirique un risque

d'erreur alpha est le risque qu'on accepte de prendre de dire que la différence n'est pas due au hasard alors qu'elle est due au hasard

- choix, a priori d'un risque « raisonnable » alpha = 5%

Autre mode de résolution pour savoir si deux paramètres différent réellement : l'approche par intervalles de confiance (IC) Calcul de précision d'une statistique

« si je multipliais les expériences donc les échantillons, 95 fois sur 100 le paramètre mesuré serait compris entre X et entre Y

C'est l'intervalle de confiance à 95% (IC95%) du paramètre mesuré (moyenne, pourcentage, risque relatif, odds ratio….)

Pour savoir si deux paramètres différent réellement on peut utiliser l'approche par intervalles de confiance.

Exemple : Proportion de guérison avec le TT A 40% ➡ IC95% = [30;50]

Proportion de guérison avec le TT B 15% ➡ IC95% [6;21]

Dans le meilleur des cas B guéris 21 % des patients

estimation la plus haute de l'intervalle de confiance

– Dans le pire des cas A guéris 30 % des patients

estimation la plus basse de l'intervalle de confiance

A semble meilleur que B

– risque alpha = 5% puisque les IC sont estimés à 95 %

Rappel IC variable quantitative

Pour la moyenne on peut donner la précision de l'estimation de la moyenne par l'intervalle de confiance de la moyenne

moyenne ± 1,96 *((écarts type) / Racine(n))

[n étant le nombre de cas]

C'est l'intervalle de confiance à 95 % de la moyenne

si l'on retirait au sort 100 fois un échantillon de même taille 95 moyennes estimées sur 100 seraient comprises dans cet intervalle

Ex : la moyenne des patients suspects d'embolie pulmonaire peut être estimée à

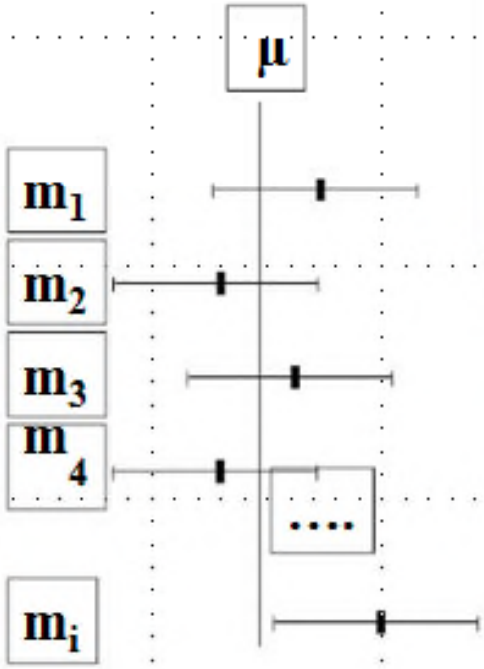
63 ans ± 1 an

Ce n'est pas la dispersion des valeurs individuelles mais

la précision de la mesure

IC d'une Moyenne

Estimation IC 95%



Attention !
μ reste constant
C'est l'intervalle de confiance qui varie autour de μ pour chaque échantillon.

Si l'on a choisi un risque de 5%, en moyenne, l'estimation obtenue dans 1 échantillon sur 20 ne contiendra pas la vraie valeur μ.

IC d'une Proportion

Un médecin observe 20 cas de guérison parmi 50 cas de cancer.

Quel est l'intervalle de confiance qui a 95 chances sur 100 de contenir la vraie valeur du pourcentage de guérison ?

$$IC = \left[0.4 \pm 1.96 * \sqrt{\frac{0.4 * 0.6}{50}} \right] = \left[0.4 \pm 1.96 * 0.07 \right] = \left[0.2628; 0.5372 \right]$$

où *1.96* est la valeur de la variable N(0,1) correspondant au risque choisi : 5 %

Estimation ponctuelle sur 50 cas : 40 % (pourcentage observé)

Intervalle de confiance : 26 % à 54 %

RESUME IMPORTANT INTERVALLE DE CONFIANCE

Construit autour de la moyenne observée sur l' échantillon

Construit en utilisant la variance observée

Définit l'intervalle « raisonnable » dans lequel la moyenne vraie (théorique) peut se situer au risque 1-α

Si deux IC95% sont disjoints alors le test statistique correspond est significatif au seuil 5%

Principe d’un test statistique

Pour comparer deux paramètres (deux moyennes par exemple) on va se ramener à une valeur qui suit une loi de distribution connue

Ex : N (0,1) pour la différence de deux moyennes

On va ensuite regarder sur une table de cette loi de distribution, si la valeur observée est une valeur « étonnante » ie peu probable pour cette loi de distribution « banale »

Objectif d’un test statistique

Un test permet de porter une affirmation en contrôlant le risque d’erreur

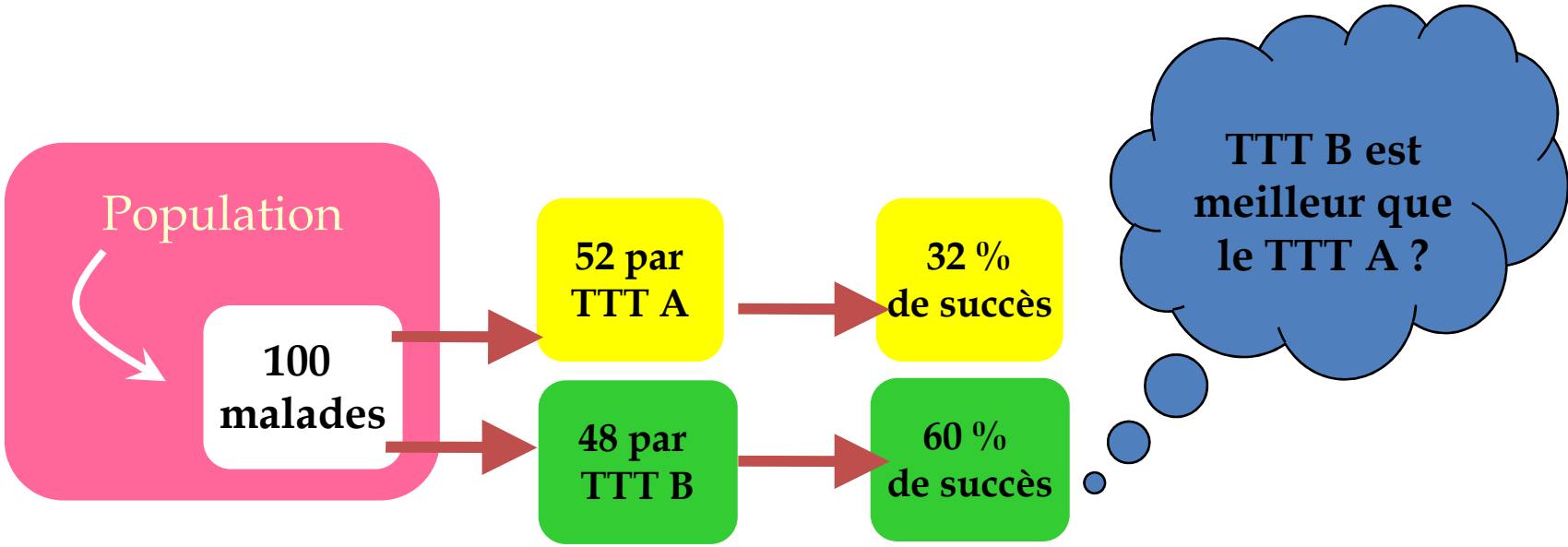
Il donne une réponse à la question :

La différence observée entre mes deux paramètre peut-elle être due aux fluctuations d’échantillonnage ?

Les deux paramètres sont-ils deux estimations d’une même population théorique ?

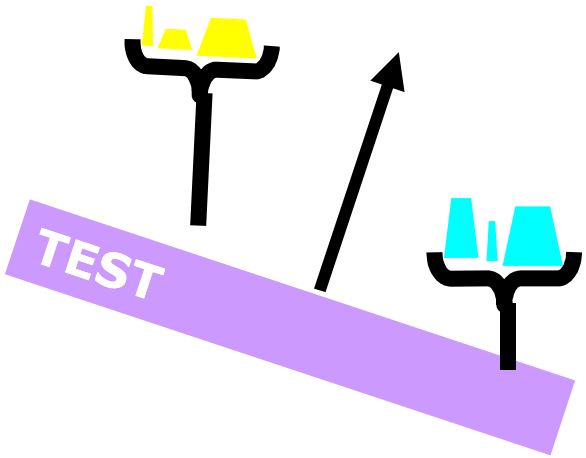
- Au moment de son examen physique annuel, Monsieur Salah présentait PA diastolique (PAD) de 97 mm Hg.
- Vous lui avez conseillé de diminuer sa consommation de sel, de perdre du poids et de faire de l'exercice.
- Bien qu'il ait fait quelques efforts dans cette direction, la PAD mesurée lors de ses trois dernières consultations était de 92, 96 et 93 mm Hg.
- Vous avez démarré un traitement médicamenteux et vous avez continué à suivre M Salah. Les mesures suivantes des PAD étaient de 88, 91 et 86 mm Hg.
- Le traitement a-t-il été efficace ?

- Quel est le meilleur traitement, A ou B ?



Analyse Univariée → **ESTIMATION**

Analyse Bivariée → **COMPARAISON**

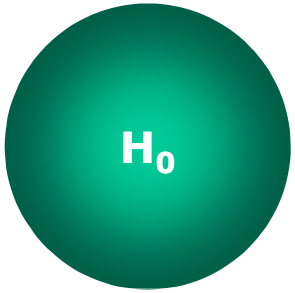


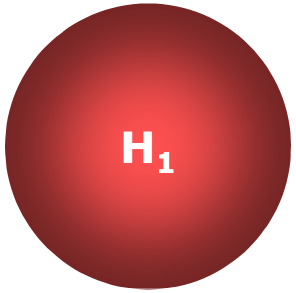
- Une comparaison porte sur des séries de données (moyennes, pourcentages, etc.)
- Test statistique = pesée
- Comparaison trouve une différence parfois grande.
- Voir si cette différence est simplement liée au hasard, ou elle est bien réelle
- Extrapoler aux populations avec un risque d'erreur

- Cette différence pourrait être expliquée par le hasard = fluctuations d'échantillonnage
- Le test statistique va permettre de quantifier le rôle du hasard dans l'observation de cette différence
- La décision est basée sur le test statistique
- La formulation de l'hypothèse nulle est la première étape du test d'hypothèse (test statistique)
- Elle peut être vérifiée
  - Rejetée
  - Gardée

Hypothèses, oui mais lesquelles ?

Nulle





Alternative

Pas de différence
Indépendance
Les paramètres d'où sont issus
les échantillons étudiés sont identiques

A = B

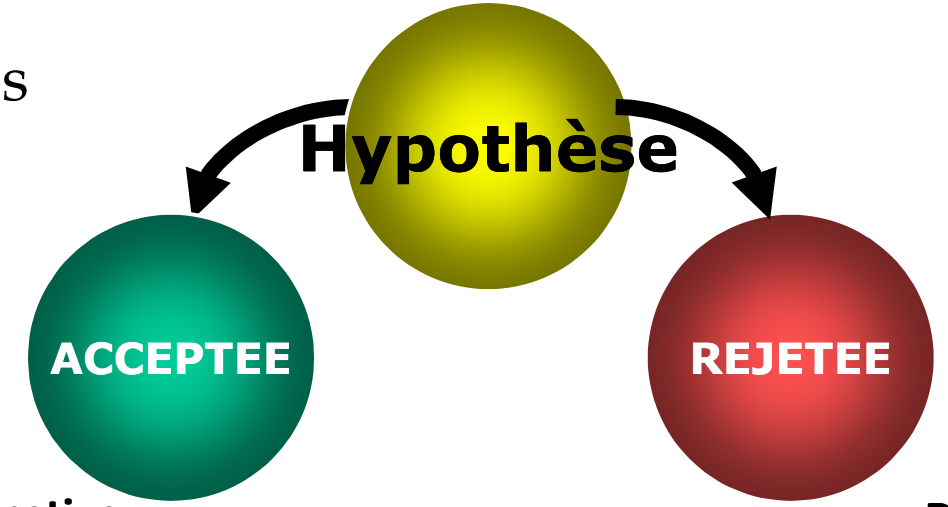
Il existe une différence
Les paramètres d'où sont
Issus les échantillons sont différents

A ≠ B
A > B
A < B

Conditions d'utilisation d'un test

- Un test n'a de sens que s'il teste une hypothèse préalablement posée

- 2 possibilités



Différence non significative

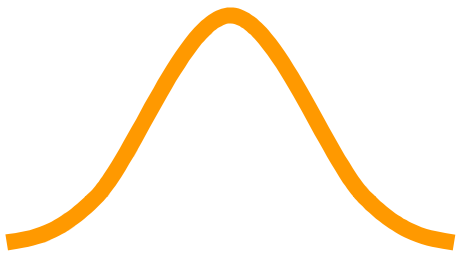
Différence significative

Contre exemple….

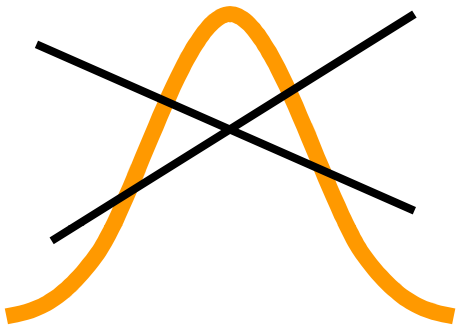
- On décide de comparer le poids des individus qui passent dimanche matin sur les trottoirs de droite et de gauche de l'avenue Didouche Mourad á Alger
- Il n'est pas impossible de trouver une différence et même parfois significative.
- Mais ceci n'aurait aucun sens et la recherche d'une explication a posteriori serait absurde

2 familles de tests.....

PARAMETRIQUE

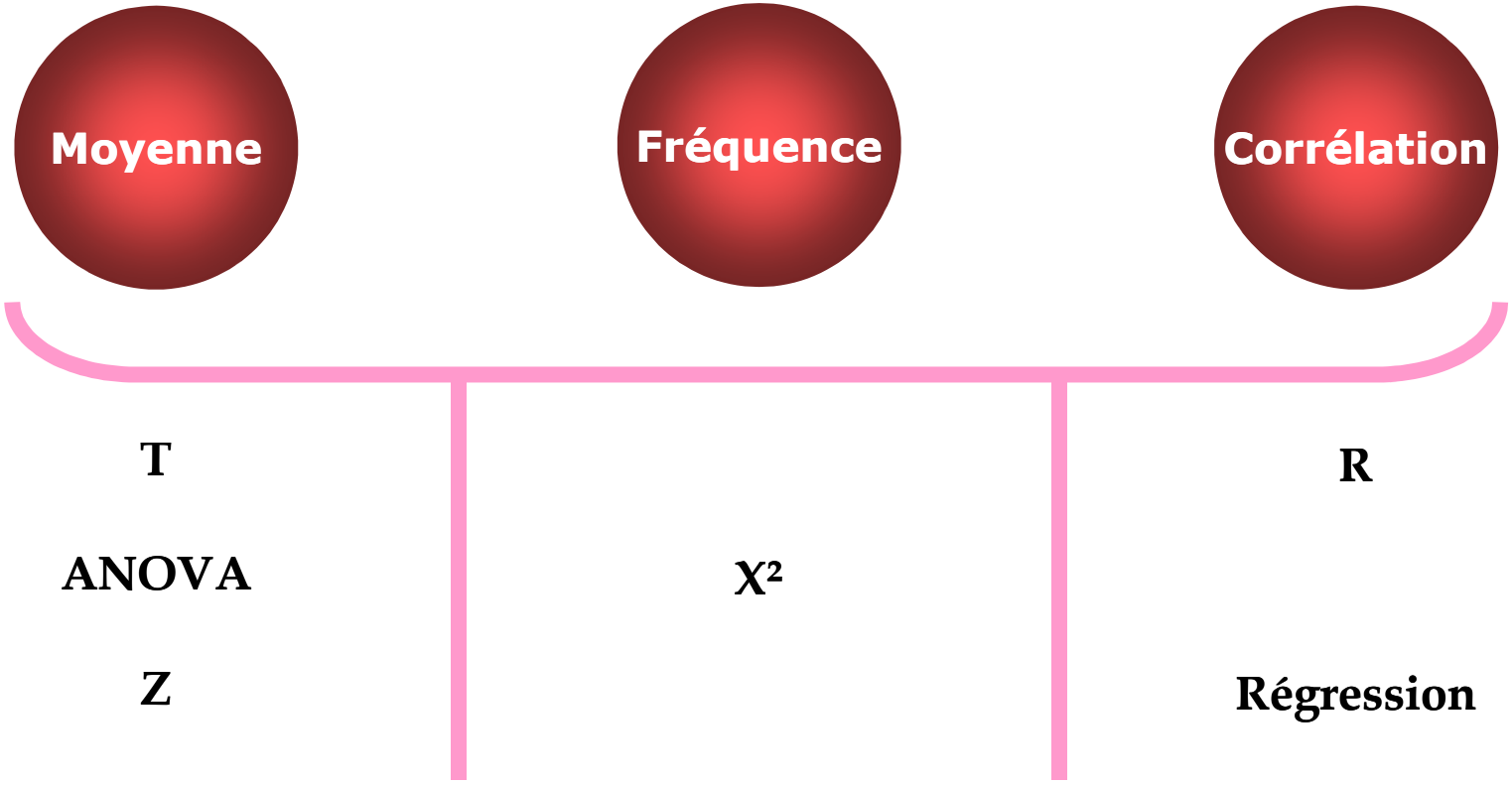


NON PARAMETRIQUE

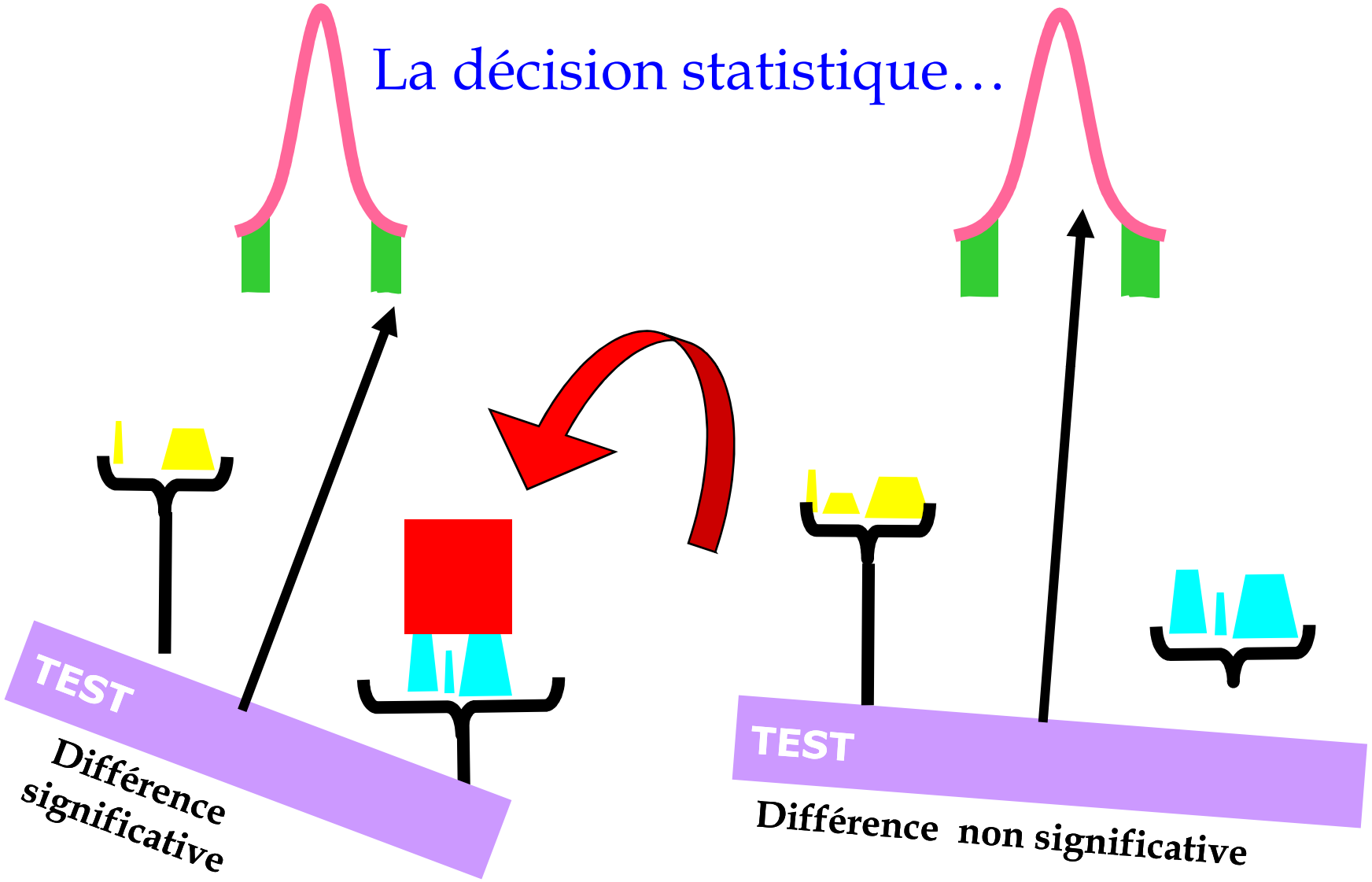


Test paramétrique	Test non paramétrique
Test t de Student	Test de Mann et Whitney
Test du Chi deux..	Test de Wilcoxon
Analyse de variance	Test de Kruskall et Wallis*
Corrélation linéaire	Test de Spearman

Tests paramétriques



La décision statistique…



Seuil de signification « p » (1)

- Le test statistique donne la probabilité « p » que le hasard puisse expliquer les résultats
- Si la probabilité « p » est inférieure ou égale à un certain seuil, appelé seuil de signification,

on rejette H0 et on dit que la différence est significative

- Si « p » est supérieure au seuil,

on ne rejette pas H0 et on dit que la différence n'est pas significative

- Le seuil de signification est généralement fixé à 5 %
- Convention très largement adaptée

Au total : les étapes d'un test statistique

Nature des variables (VQN, VQL), distribution normale ?

Choix du test statistique

Définir Hypothèse nulle et alternative (H0 et H1)

Fixer le seuil de signification α et se rappeler du caractère antagoniste de β

Mécanique du calcul (indiquer le test et le calculer)

Rejeter H_0 ou pas! Et décision…

Test de Shapiro et Wilk
Test de normalité

Ce test vérifie, si une série se distribue de façon normale

Démarche de vérification :

1/ Classer les différentes valeurs de la série par ordre croissant

Exemple : soit le tableau suivant correspondant aux résultats de mesures d'un alésage (en mm)

Pièce n°	1	2	3	4	5	6	7	8	9	10
mesure	12.124	12.230	12.327	12.242	12.466	12.215	12.026	12.359	12.215	12.387

1. Classement des valeurs de mesure par ordre croissant :

12.026	12.124	12.215	12.215	12.230	12.242	12.327	12.359	12.387	12.466

2/ Calculer la moyenne X de la série de mesure : $\overline{x} = 12.259$

3/ Calculer la variance de la série de mesure : $\delta^2 = 0.1514$

4/ Calculer les différences respectives :

$$d_1 = x_n - x_1$$

$$d_2 = x_{n-1} - x_2$$

6 / Calculer la valeur :

Dans notre exemple :

$$b = \sum a_i d_j = 0.2525 + 0.0865 + 0.0308 + 0.0137 + 0.0005 = 0.384$$

7 / Calculer le rapport :

Dans notre exemple :

$$w = \frac{b^2}{\delta^2} = \frac{(0.384)^2}{0.1514} = 0.9739$$

8 / Comparer W calculé au $_{Wcritique}$ de la table, avec n nombre de données.

Si W calculé est supérieur au $_{Wcritique}$ de la table, la normalité est acceptée.

Si W calculé est inférieur au $_{Wcritique}$ de la table, la normalité est rejetée

Dans le cas de l'exemple, W = 0.9739 > 0.842

l'hypothèse de normalité est acceptée.

(Si W < 0.842, il y aurait refus avec un risque de 5% de rejeter une distribution normale.)

n	2	3	4	5	6	7	8	9	10	
J										
1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739	
2		0.0000	0.1677	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291	
3				0.0000	0.0875	0.1401	0.1743	0.1976	0.2141	
4						0.0000	0.0561	0.0947	0.1224	
5								0.0000	0.0399	

n	11	12	13	14	15	16	17	18	19	20
J										
1	0.5601	0.5475	0.5359	0.5251	0.5150	0.5056	0.4963	0.4886	0.4808	0.4734
2	0.3315	0.3325	0.3325	0.3318	0.3306	0.3290	0.3273	0.3253	0.3232	0.3211
3	0.2260	0.2347	0.2412	0.2460	0.2495	0.2521	0.2540	0.2553	0.2561	0.2565
4	0.1429	0.1586	0.1707	0.1802	0.1878	0.1939	0.1988	0.2027	0.2059	0.2085
5	0.0695	0.0922	0.1099	0.1240	0.1353	0.1447	0.1524	0.1587	0.1641	0.1686
6	0.0000	0.0303	0.0539	0.0727	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334
7			0.0000	0.0240	0.0433	0.0593	0.0725	0.0837	0.0932	0.1013
8					0.0000	0.0196	0.0359	0.0496	0.0612	0.0711
9							0.0000	0.0163	0.0303	0.0422
10									0.0000	0.0140

N	W ‘95%’	W ‘99%’
10	0.842	0.781
11	0.850	0.792
12	0.859	0.805
13	0.856	0.814
14	0.874	0.825
15	0.881	0.835
16	0.837	0.844
17	0.892	0.851
18	0.897	0.858
19	0.901	0.863
20	0.905	0.868
21	0.908	0.873
22	0.911	0.878
23	0.914	0.881
24	0.916	0.884
25	0.918	0.888
26	0.920	0.891
27	0.923	0.894
28	0.924	0.896
29	0.926	0.898
30	0.927	0.900
31	0.929	0.902
32	0.930	0.904

TESTS PARAMETRIQUES

TEST DE STUDENT

Ce test permet de comparer deux distributions extraites d'une population normale ou approximativement normale au niveau de leurs moyennes.

Il s'agit de décider si la différence observée entre les moyennes des deux échantillons de comparaison est attribuable à la variable indépendante testée ou si elle peut être considérée comme l'effet du hasard.

1/ Cas de deux échantillons indépendants

Il s'agit de deux variables:

VI = variable indépendante: nominale dichotomique / VD = Variable dépendante: variable d'intervalle

séries de mesure pour lesquelles il n'y a aucune correspondance entre les éléments de la première série et ceux de la deuxième; les deux séries de mesures sont obtenues avec des sujets différents. Dans ce cas le but de l'application du test t est de voir si les deux moyennes calculées sur les deux échantillons diffèrent significativement.

Soit la situation suivante :

Variable

HYPOTHSE

H0: m1 = m2 (c'est-à-dire les deux groupes de comparaison appartiennent à des populations qui possèdent des moyennes identiques)

H1: $m1 \neq m2$ () ou m1 < m2 ou m1 > m2 (Hypothèses unilatérales)

hypothèse bilatérale

Conditions d'application

La distribution des données de chaque échantillon ne peut pas différer fortement de la normale,

et, en particulier, ne pas être trop dissymétrique, surtout si les échantillons sont petits

- Les variances des populations de provenance ne peuvent pas être extrêmement différentes

- Les tailles des échantillons ne peuvent pas extrêmement différentes

$$t = \frac{m_1 - m_2}{\sqrt{V_C\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

m_1 : moyenne du premier échantillon

m_2 : moyenne du deuxième échantillon

n_1 : nombre de mesures (sujets) du premier échantillon

n_2 : nombre de mesures (sujets) du deuxième échantillon

Vc: la variance commune, c'est une sorte de moyenne des deux variances (V1 et V2) pondérée par le nombre de mesures n1 et n2; sa formule est:

$$n_1 \neq n_2 \Rightarrow V_C = \frac{V_1(n_1 - 1) + V_2(n_2 - 1)}{n_1 + n_2 - 2}$$

$$n_1 = n_2 \Rightarrow V_C = \frac{V_1 + V_2}{2}$$

Une fois on a la valeur de t calculé, se rapporter à la table de t de Student pour comparer le "t calculé" au "t critique" , et ce, au ddl = n1 + n2 -2 et au seuil 0,05.

La différence est significative si "t calculé" est supérieur ou égal au "t critique".

Vérification de la normalité des distributions

$$CD_1 = \frac{3(\mu_1 - M_d)}{\partial_1} = \frac{3(63.5 - 63)}{15.6} = 0.096$$

$$CD_2 = \frac{3(\mu_2 - M_d)}{\partial_2} = \frac{3(48.7 - 49)}{16.4} = -0.054$$

Vérification de l'homogénéité des variances

$$F = \frac{(16.4)^2}{(15.6)^2} = 1.105$$

La valeur critique de F à ddl horizontal = 17 (18-1) et ddl vertical = 26 (27-1) égal 1,89

(17 et 26 ne figurent pas sur la table, on prendra les valeurs immédiatement supérieures.

F calculé étant inférieur à F critique, on conclue donc que la différence entre les variance n'est pas significative

$$V_C = \frac{(15.6)^2(27-1)+(16.4)^2(18-1)}{27+18-2} = 253.48$$

$$t = \frac{63.5 - 48.7}{\sqrt{253.48\left(\frac{1}{27}+\frac{1}{18}\right)}} = 3.06$$

Nous devons maintenant chercher la valeur critique de t.

ddl = 27+18-2 = 43; au seuil 0,05 t = 2,02 (43 n'existe pas sur la table, on choisira le degré de liberté juste inférieur c'est-à-dire 40).

3,06 étant > 2,02,

nous rejetons H0 et nous admettons l'existence d'une différence significative entre m1 et m2.

Cas de deux échantillons dépendants ou appareillés

Il s'agit de deux séries de mesures pour lesquelles il y a une correspondance stricte, terme à terme, entre les éléments de l'une et les éléments de l'autre. C'est le cas par exemple de deux séries de notes. relevées auprès d'un échantillon d'élèves, la première avant les vacances et la deuxième à la rentrée; il y a donc une correspondance parfaite puisque c'est le même groupe qui effectue les deux épreuves.

Là encore on va calculer un t qui indique si les deux moyennes sont significativement différentes. La formule sera légèrement modifiée par rapport à la précédente:

$$t = \frac{m_d}{\partial_d / \sqrt{N}} = \frac{\sum |m_1 - m_2|}{\partial_d / \sqrt{N}}$$

md: la moyenne des différences
σd: l'écart-type de la distribution des différences
N: le nombre de sujets

Exemple

dans une expérience sur la perception du langage, on fait subir deux épreuves à un même groupe de 40 sujets. On a obtenu les mesures suivantes qui désignent le nombre de mots correctement reproduits:

Sujets	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Epreuve1	3	5	5	7	7	7	4	6	6	7	4	8	5	8	6	8	6	7	6	7
Epreuve2	5	2	4	2	6	3	4	1	3	4	1	3	3	2	5	3	2	7	3	3

Sujets	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
Epreuve1	9	7	6	5	7	5	7	9	6	7	6	6	8	4	6	4	8	5	5	8
Epreuve2	3	3	5	2	4	2	6	3	2	4	3	5	2	4	2	4	5	3	4	4

On commence par calculer les différences entre les mesures de l'épreuve1 et celles de l'épreuve2
puis les relever au carré (d^2). On obtient la distribution suivante:

on calcule la variance puis l'écart-type de cette distribution des différences (Vd et σd)

$$V_d = \frac{\sum d^2 - \frac{(\sum d)^2}{N}}{N-1} = \frac{474 - \frac{(114)^2}{40}}{39} = 3.82$$

L'écart-type des différences

$$\partial_d = \sqrt{V_d} = \sqrt{3.82} = 1.95$$

4- On applique la formule de t pour échantillons appareillés

$$t = \frac{m_d}{\partial_d / \sqrt{N}} = \frac{2.85}{1.95 / \sqrt{40}} = 9.23$$

dans la table de t, on cherche la valeur critique au ddl N-1 = 40-1 =39 et au seuil 0,05; on trouve t= 2,02 ce qui est largement inférieur à t calculé,

la différence entre les deux moyennes est donc très significative

Test de Khi-deux (X^2)

Il permet de comparer deux ou plusieurs groupes caractérisés par une répartition de leurs effectifs respectifs.

1) Cas des échantillons indépendants

1. Ce test n'est applicable que si les catégories sont les mêmes dans les différents échantillons

2. Les données doivent être indépendantes d'une colonne à l'autre ou d'une rangée à l'autre (pas d'échantillons appareillés).

3. Les groupes doivent avoir une taille suffisante, ce test ne pas être appliqué si 20% ou plus des fréquences attendues sont inférieures à 5, sinon il faut apporter la correction de Yates.

Test de Khi-deux (X^2)

Calculer l'effectif théorique pour chaque case

Calculer la statistique khi-deux pour chaque case

Faire la somme des khi-deux obtenus

Comparer ce résultat avec la valeur tabulaire correspondant au seuil de signification choisi et au nombre de degré de liberté que comporte la situation. Si le résultat est supérieur ou égal à cette valeur, alors on rejette H0

Soit une variable nominale trichotomique VA formée de 2 modalités: a1 et a2

Soit une variable ordinale de catégories rangées VB à 3 modalités: b1; b2 et b3

1/Dresser le tableau des effectifs observés

	b1	**b2**	**b3**	**Total**
a1	n1	n2	n3	L1
a2	n4	n5	n6	L2
Total	C1	C2	C3	**N**

Test de Khi-deux (X^2)

Calculer les effectifs théoriques (appelés également attendus)

	b1	b2	b3	Total
a1	n'1	n'2	n'3	L1
a2	n'4	n'5	n'6	L2
Total	C1	C2	C3	**N**

L:Total ligne

C:Total colonne

N: Effectif total

	b1	b2	b3
a1	n'1=$C1 \times L1/N$	n'2= $C2 \times L1 / N$	n'3= $C3 \times L1 / N$
a2	n'4= $C1 \times L2/N$	n'5=$C2 \times L2/N$	n'6= $C3 \times L2/N$

Calculer le Khi-deux des cases

Pour chaque case, on applique: (effectif observé – effectif théorique)² /effectif théorique

	b1	b2	b3
a1	$(n1-n'1)^2/n'1$	$(n2-n'2)^2/n'2$	$(n3-n'3)^2/n'3$
a2	$(n4-n'4)^2/n'4$	$(n5-n'5)^2/n'5$	$(n6-n'6)^2/n'6$

Test de Khi-deux (X^2)

Si 20% au plus des effectifs théoriques sont inférieurs à 5, on apporte la correction de Yates et la formule devient: (|effectif observé – effectif théorique| - 0.5) / effectif théorique

Calculer le Khi-deux (la somme de chacune des cases de l'étape précédente).

Déterminer les degrés de liberté de Khi-deux en appliquant la formule suivante:

ddl=(Nombre de colonnes – 1) (Nombre de lignes – 1)

La valeur de Khi-deux calculée doit être comparée à la valeur critique de Khi-deux (sur la table)

au seuil 0,05: si Khi-deux calculé est supérieur au Khi-deux théorique, on considère la différence significative, c'est-à-dire qu'il y a influence de la variable indépendante sur la variable dépendante.

Remarque: cette procédure est générale, qu'il s'agisse d'un tableau à quatre cases ou plus.

Test de Khi-deux (X^2)

Khi-2 des cases

Puisque 12 fréquences théoriques sur 16 sont inférieures à 5, on applique la correction de Yates on obtient le tableau suivant:

$$\chi^2_{yates} = \sum^{k} \frac{(|f_0 - f_{th}| - 0.5)^2}{f_{th}}$$

	Maths	Lettres	Sciences	Technique
Très défavorisée	0.10	0.01	1.04	0.56
Défavorisée	0.009	0.05	0.04	0.61
Moyenne	0.11	0.05	0.04	0.0025
Favorisée	0.008	0.3	1.32	0.002

$$\chi^2_{cal} = 4.25$$

Test de Khi-deux (X^2)

$$ddl = (4-1)*(4-1) = 9$$

$$\chi^2_{critique} = 16.9$$

$$\alpha = 0.05$$

$4.25 < 16.9$, donc H0 est retenue.

La catégorie socio professionnelle n'a pas d'effet sur le choix de la filière.

Test de Khi-deux (X^2)

Cas des échantillons dépendants

Il s'agit de comparer un tableau de fréquences construit sur des dichotomies (fréquences recueillies auprès d'un seul échantillon à des moments différents ou dans deux situations différentes).

Supposant, par exemple que l'on veuille étudier la différence entre le nombre d'élèves accédant à deux types de formation

FORMATION B		FORMATION A	
	ADMIS	REFUSES	TOTAL
ADMIS	n1	n2	N1
REFUSES	n3	n4	N2
TOTAL	N3	N4	N

Ce sont les mêmes candidats (ayant participé à l'examen de la formation A et l'examen de la formation B),

on veut comparer la proportion des admis à la première formation avec la proportion des admis à la deuxième formation)

c'est - à - dire les fréquences :

$$P_1 = \frac{N_3}{N} \text{ et } P_2 = \frac{N_1}{N}$$

Pour ce faire, on calcule un χ^2 assez différent du précédent : $\chi^2 = \frac{(n_2 - n_3)^2}{n_2 + n_3}$

Remarquons que cette formule ne s'intéresse qu'aux effectifs des cases hétérogènes (admis à une formation et refusés à une autre).

Exemple (tiré de S. Ehrlich et C. Flament, 1970, p.158):

On a posé à 300 personnes deux questions: "allez-vous souvent au cinéma?" et "allez-vous souvent au théâtre?".

Les réponses sont "oui" ou "non". On observe les résultats suivants:

		CINEMA		
		OUI	NON	TOTAL
THEATRE	OUI	n1=42	n2=48	N1=90
	NON	n3=78	n4=132	N2=210
	TOTAL	N3=120	N4=180	N=300

42 personnes vont souvent au **cinéma** et au **théâtre**;

78 personnes vont souvent au cinéma et rarement au théâtre;

120 personnes vont souvent au cinéma;

90 personnes vont souvent au théâtre

La question: la différence entre ces deux nombres est-elle significative?

on calcule un χ^2 assez différent du précédent :

$$\chi_{Cal}^{\ 2} = \frac{(n_2 - n_3)^2}{n_2 + n_3} = \frac{(48-78)^2}{45+78} = \frac{900}{126} = 7.14$$

La valeur critique $\chi_{Cri}^{\ 2}$ pour $\text{ddl} = 1$ et 0,05 probabilité d'erreur donne 3.84

la différence est dons significative.

les tests non paramétriques

Il existe de tests moins "exigeants" en conditions d'applications, notamment en ce qui concerne la taille de l'échantillon, la normalité de la distribution et l'égalité des variances, ces tests sont dits non paramétriques.

Le principe de base de ces tests est de transformer les données en rangs et à mesurer

l'accord entre les rangs observés et ce que devrait être ces rangs sous une hypothèse nulle. Parmi ces tests nous allons voir:

le test de Mann-Wihtney, l'alternative non paramétrique de t de Student pour deux échantillons indépendants;

le test de Wilcoxon, l'alternative non paramétrique de t de Student pour deux échantillons dépendants;

le test de Kruskal-Wallis, l'alternative non paramétrique de l'analyse de variance.

U de Mann-Withney

Ce test est destiné à étudier si une variable indépendante nominale dichotomique influence une variable dépendante ordinale de scores rangés ou d'intervalle.

Ce test doit être préféré au test t de student lorsque la distribution n'obéit pas à la loi normale (donc remarquablement dissymétrique)

U de Mann-Withney

Algorithme de résolution

cas : n_A et n_B sont supérieurs à 8

Supposant les donnés suivantes:

A	11	9	7	12	12	40	5	4	15	10	10	14
B	13	15	15	14	35	18	13	25	20	6	5	

$n_A = 12$

$n_B = 11$

Transformer les scores en rangs

Mélanger les donnés de deux groupes

Ordonner la série obtenue en ordre croissant

Accorder des rangs; pour les ex-æquo attribuer à chacun le rang moyen

Reconstruire les deux groupes avec données et les rangs correspondants

U de Mann-Withney

calculer U et U'

$$U = n_A n_B + \frac{n_A(n_A+1)}{2} - T_A$$

$$U' = n_A n_B + \frac{n_B(n_B+1)}{2} - T_B$$

Mann et Whitney ont montré que la variable U se distribue selon une loi approximativement normale.

Calculer donc: la moyenne et l'écart-type de la distribution de U:

$$m_U = \frac{n_A * n_B}{2} \qquad \delta_U = \sqrt{\frac{(n_A * n_B)(n_A + n_B + 1)}{12}}$$

Il en est de même pour U', valeur symétrique de U. il suffit donc de tester l'écart entre U et m (ou entre m et U')

$$|Z| = \frac{|U - m_U|}{\delta_U}$$

Si nous revenons à notre exemple, nous aurons donc:

$$U = n_A n_B + \frac{n_A(n_A+1)}{2} - T_A = (12*11) + \frac{12*13}{2} - 114 = 96$$

$$\mathrm{U}' = n_A n_B + \frac{n_B(n_B+1)}{2} - T_B = (12*11) + \frac{11*10}{2} - 162 = 36$$

$$m_U = \frac{12*11}{2} = 66 \quad \delta_{\mathrm{U}} = \sqrt{\frac{(12*11)(12+11+1)}{12}} = 16.25$$

On peut vérifier que : $\frac{\mathrm{U}+\mathrm{U}'}{2} = \frac{96+66}{2} = 66 = m_U$

Par conséquent : $|\mathrm{Z}| = \frac{|96-66|}{16.25} = 1.85$

Vérifier la signification de la valeur Z :

Si Z calculé est supérieur ou égal à 1.96, la différence est significative au $\mathrm{P} = 0.05$

Si Z calculé est supérieur ou égal à 2.56, la différence est significative au $\mathrm{P} = 0.01$

Algorithme de résolution

$$\text{cas : } n_A \text{ et } n_B \text{ sont inférieurs à } 8$$

Dans ce cas, la distribution n'est pas gaussienne, le modèle précédent ne peut pas être appliqué.

Mann et Withney ont construit des tables avec des valeurs critiques qu'il est possible de consulter directement en fonction de:

de **U** si U est inférieur à U'

de **U'** si U' est inférieur à U

Supposons les mesures de deux groupes et leurs rangs:

A	Scores	6	3	10	5	14	$n_A = 5$
	Rangs	7	9	5	**8**	**3**	$T_A = 32$
B	Scores	12	8	16	18		$n_B = 4$
	Rangs	**4**	6	2	**1**		$T_B = 13$

$$U = 5*4 + \frac{5(5+1)}{2} - 32 = 3$$

$$U' = 5*4 + \frac{4(4+1)}{2} - 13 = 17$$

La table est consultée en fonction de l'effectif n2 du plus grand de deux échantillons (ici, n2 =5).

Pour n1=4 et U=3, nous lisons P=.056

L'hypothèse nulle n'est pas rejetée, il n'existe pas une différence entre les moyennes des rangs.

Remarque:

Pour simplifier les calculs on prendra toujours la somme des rangs dans la situation comportant le moins de sujets.

Lorsqu'il y aura le même nombre de sujets dans les deux conditions, il sera possible de prendre l'une ou l'autre des deux conditions pour calculer la somme des rangs.

Le test de Kruskal-Wallis

C'est la généralisation du test de Mann-Withney à trois échantillons ou plus . Les scores sont remplacés par les rangs obtenus à l'intérieur d'un seul groupe constitués à partir des échantillons à comparer.

Supposant 4groupes de sujets reçoivent un enseignement selon quatre méthodes différentes.

On souhaite comparer leurs résultats sur la base des données suivantes:

Groupes	1	2	3	4
Scores	8	15	18	4
	20	14	16	7
	13	7	15	12
	14	9	19	10
	17	12		8
		10		6
				11
Effectifs	$n_1=5$	$n_2=6$	$n_3=4$	$n_4=7$

On mélange les 4 groupes (k=4) et on ordonne les scores: 4 - 6 - 7 - 7 - 8 - 8 - 9 - 10 - 10 - 11 - 12 - 12 - 13 - 14 - 14 - 15 - 15 - 16 - 17 - 18 - 19 - 20

On applique la formule de Kruskal et Wallis:

$$H = \left[\frac{12}{N(N+1)} \times \frac{\sum T_i^2}{n_i} \right] - 3(N+1)$$

$$H = \frac{12}{22(22+1)} \times \left[\frac{(74)^2}{5} + \frac{(61.5)^2}{6} + \frac{(75.5)^2}{4} + \frac{(42)^2}{7} \right] - 3(22+1) = 11.64$$

Cette variable H suit une loi de X^2

. Il suffit donc de revenir à la table de X^2 et de comparer H calculé à la valeur critique de X^2 au ddl = k – 1 (c'est-à-dire nombre de groupes – 1).

K = 4; k – 1= 3. La valeur critique de X^2 au ddl = 3 et P=0,05 est égale à 7,82.

La valeur calculée est supérieure à la valeur théorique, on rejette donc H0.

Autrement dit il existe des différences entre

les moyennes des rangs des 4 groupes.

Le test de Wilcoxon

Il permet la comparaison les moyennes des rangs de deux échantillons appareillés.

Son principe consiste à classer les sujets dans l'ordre croissant des valeurs absolues des différences non nulles.

Supposons les données suivantes portant sur les notes (variant de 0 à 10) obtenues par un groupe d'élèves à deux moments différents de l'année scolaire:

Eléves	A	b	c	d	e	f	g	h	i	j
Première note(A)	1	1	2	2	7	2	3	6	4	5
Deuxième note(B)	9	6	9	2	5	7	8	8	7	4

calculer pour chaque élève la différence entre la première et la deuxième note

$$D=B-A$$

ce qui donnera la distribution suivante

Eléves	a	b	c	d	e	f	g	h	i	j
D	+8	+5	+7	0	-2	+5	+5	+2	+3	-1

Nous constatons que:

Un élève n'a ni régressé ni progressé (d=0)

Deux élèves ont régressé (d négative)

Sept élèves ont progressé (d positive)

Dans notre exemple T+ est la somme des différences des sujets **a, b, c, f, g, h, i**.

T- est la somme des différences des sujets **e , j**

T+=41.5 T-=3.5

$$T^{+} = 41.5 \qquad T^{-} = 3.5$$

$$\text{Notons que} : T^{+} + T^{-} = \frac{n(n+1)}{2}$$

$$41.5 + 3.5 = \frac{9(9+1)}{2} = 45$$

tester la plus petite valeur des T+ et T-

Cas où le nombre de couples dont les différences non nulles est inférieur ou égal à 20 (n≤20)

Cas où le nombre de couples dont les différences non nulles est inférieur ou égal à 20 ($n \leq 20$)

La distribution de T n'est pas normale; la valeur théorique de T est tabulée

on rejette H0 si T Calculé (T+ ou T-) est inférieur à T lu sur la table.

Dans notre exemple, n = 9 et T calculé = 3,5 (nous avons pris T- parce qu'elle est plus petite que T+); à P = 0,05 T théorique = 6.

L'hypothèse nulle est alors rejetée.

Cas où le nombre de couples dont les différences non nulles est supérieur à 20 (n>20)

La distribution de T tend vers une distribution normale.

$$\text{Sa moyenne est : } m_T = \frac{n(n+1)}{4}$$

$$\text{Son écart - type est : } \delta_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

$$\text{On calcule alors } |Z| = \frac{|T - m_T|}{\delta_T}$$

La valeur calculée sera comparée à la valeur critique de Z (table de la loi normale réduite).

L'hypothèse nulle est rejetée si la valeur calculée de Z est supérieur à la valeur lue à un seuil donné.

Exemple:

Les données suivantes représentent les notes obtenues par un groupe d'élèves avant et après les vacances. On voulait vérifier l'hypothèse d'une déperdition des acquis:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
Avant	10	12	12	19	5	13	20	8	12	10	8	19	5	11	8	7	4	7	16	2	5
Après	8	10	8	18	8	7	12	10	7	10	3	12	8	11	5	3	5	7	9	8	14

- On commence par calculer D = B-A, ce qui donnera:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
d	2	2	4	1	-3	6	8	-2	5	0	5	7	-3	0	3	4	-1	0	7	-6	-9

Sa moyenne est : $m_T = \frac{n(n+1)}{4} = \frac{18\times19}{4} = 85.5$

Son écart - type est : $\delta_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{(18\times19)(36+1)}{24}} = 22.961$

On calcule alors $|Z| = \frac{|T - m_T|}{\delta_T} = \frac{|51-85.5|}{22.961} = 1.502$

n compare la valeur calculée de Z à la valeur lue sur la table de la loi normale réduite; Z lue au P=0,05 égale 0,121,

on rejette alors l'hypothèse nulle;

il y a une différence entre les scores des élèves avant les vacances et ceux d'après.

ANALYSE DE LA VARIANCE

L'analyse de la variance (ANOVA) a pour objectif d'étudier l'influence d'un ou plusieurs facteurs sur une variable quantitative.

Nous nous intéresserons ici au cas où les niveaux, ou modalités, des facteurs sont fixés par l'expérimentateur. On parle alors de modèle fixe.

C'est la comparaison de moyennes pour plusieurs groupes (> 2).

Il s'agit de comparer la variance intergroupe (entre les différents groupes) à la variance intragroupe (somme des fluctuations dans chaque groupe).

S'il n'y a pas de différence entre les groupes, ces deux variances sont (à peu près) égales. Sinon, la variance intergroupe est nécessairement la plus grande.

L'ANOVA se résume à une comparaison multiple de moyennes de différents échantillons constitués par les différentes modalités des facteurs. Les conditions d'application du test paramétrique de comparaison de moyennes s'appliquent donc à nouveau.

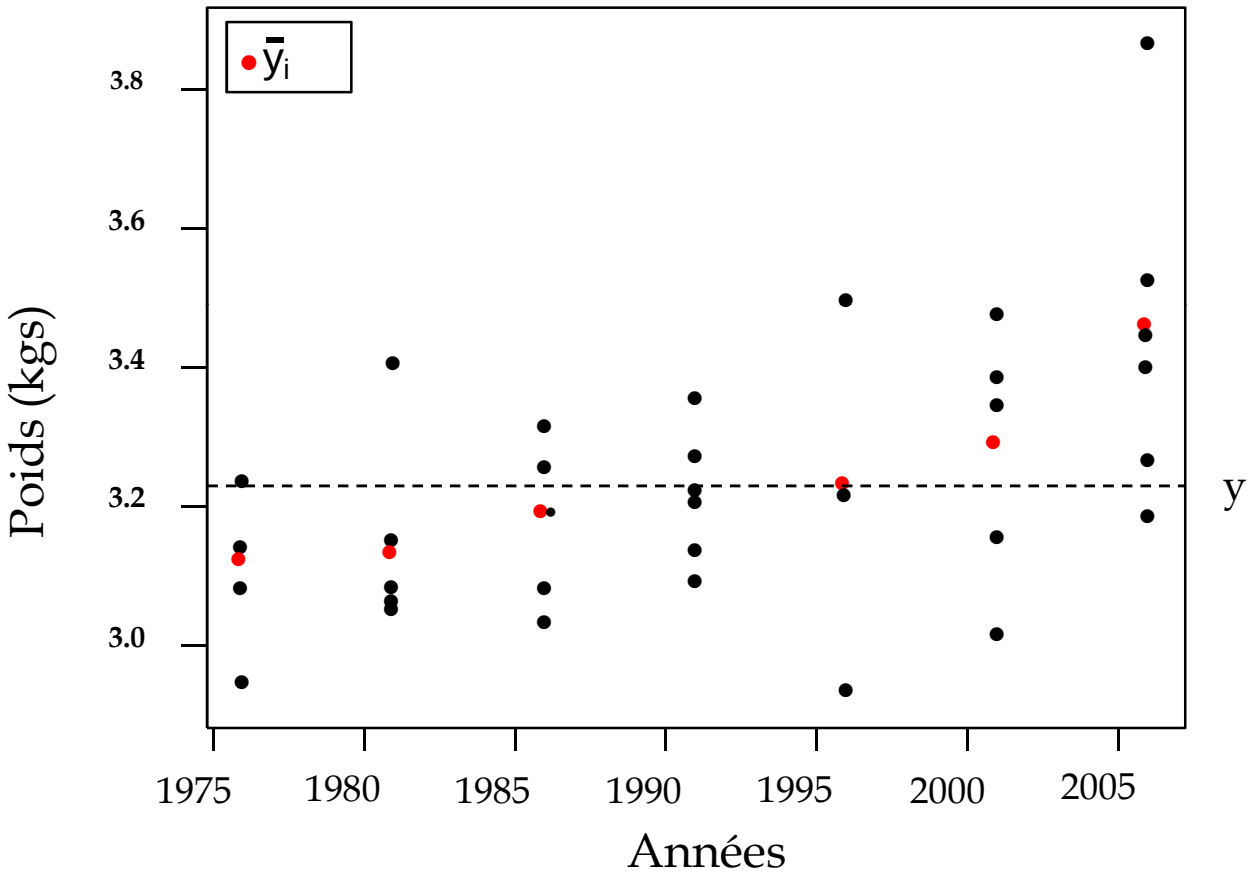
L'analyse de variance (analysis of variance ou ANOVA) peut être vue comme une généralisation du test de Student.

Evolution du poids des marmottes

On dispose des poids de différents individus, au total $N = 37$, pour chaque année.

Année	1976	1981	1986	1991	1996	2001	2006
Poids (kgs)	2.95	2.99	3.07	3.11	2.94	3.34	3.87
	3.24	3.00	3.26	3.26	3.18	3.02	3.41
	3.12	3.41	3.19	3.30	3.50	3.16	3.27
	3.05	3.02	3.32	3.14	3.22	3.48	3.37
		3.05	3.11	3.21		3.39	3.19
		3.13		3.36		3.35	3.53

Anova à un facteur



Anova à un facteur

$$SCE_{totale} = (x_1 - \bar{x})^2 + (x_1 - \bar{x})^2 + ... + (x_i - \bar{x})^2$$

$$SCE_{totale} = (2.95 - 3.23)^2 + (3.24 - 3.23)^2 + ... + (2.99 - 3.23)^2$$

$$SCE_{Inter} = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + ... + n_i(\bar{x}_i - \bar{x})^2$$

$$SCE_{Inter} = 4(3.09 - 3.23)^2 + 6(3.1 - 3.23)^2 + ... + 6(3.44 - 3.23)^2$$

$$SCE_{Intra} = SCE_{totale} - SCE_{Inter}$$

Année	1976	1981	1986	1991	1996	2001	2006
Poids (kgs)	2.95	2.99	3.07	3.11	2.94	3.34	3.87
	3.24	3.00	3.26	3.26	3.18	3.02	3.41
	3.12	3.41	3.19	3.30	3.50	3.16	3.27
	3.05	3.02	3.32	3.14	3.22	3.48	3.37
		3.05	3.11	3.21		3.39	3.19
		3.13		3.36		3.35	3.53
$\bar{x}_i$ (kgs)	3.09	3.1	3.19	3.23	3.21	3.29	3.44
n_i	4	6	5	6	4	6	6

Anova à un facteur

$$ddl_{tot} = N - 1 \qquad ddl_{\text{int } er} = nbr_{col} - 1 \qquad ddl_{\text{int } ra} = ddl_{tot} - ddl_{\text{int } er}$$

$$ddl_{tot} = 37 - 1 = 36 \qquad ddl_{\text{int } er} = 7 - 1 = 6 \qquad ddl_{\text{int } ra} = 36 - 6 = 30$$

$$CM = \frac{SCE}{ddl} \qquad F_{Cal} = \frac{CM_{\text{int } er}}{CM_{\text{int } ra}}$$

Variabilité	ddl	SCE	CM	F_{Cal}	$F_{critique}\,(a = 0.05)$
Totale	36	1.327			
Inter	6	0. 476	0.079		
Intra	30	0.851	0.028	2.821	2.42

On peut donc conclure, au risque de 5%, que le facteur temps a bien un effet sur le poids des marmottes.

Anova à deux facteurs

Les résultats d'une analyse de la variance à deux facteurs avec répétitions sont habituellement présentés dans un tableau comme celui-ci

Analyse de variance à deux facteurs avec répétitions

Source	Somme des carrés	ddl	Moyennes des carres	Fcal
Facteur A	SCFA	I-1	MCFA	MCFA / MCE
Facteur B	SCFB	J-1	MCFB	MCFB / MCE
Interaction	SCAB	(I-1)(J-1)	MCI	MCI / MCE
Résidus	SCR	IJ(K-1)	MCE	
Totale	STC	IJK-1		

Comparaison de trois types d'irrigation

Un agriculteur veut savoir les effets de trois types du système d'irrigation (R1-R2-R3) dans deux type de sols différents (S1-S2).

A partir des échantillons prélèves au nombre de quatre mesure d'humidité dans chaque type de sol (k=1,….4) associe a un type d'irrigation

	Répétition	R1 (j=1)	R2 (j=2)	R3 (j=3)
S1 (i=1)	K=1	43	41	42
	K=2	45	42	44
	K=3	46	43	46
	K=4	53	44	48
S2 (i=2)	K=1	40	35	37
	K=2	40	37	39
	K=3	40	40	40
	K=4	43	40	40

Notre objectif est de tester l'hypothèse d'égalité des moyennes des six échantillons associés à deux facteur (type de sol et type d'irrigation)

$$\bar{x} = \frac{(43 + 45 + ... + 40)}{24} = 42$$

Répétition	$n = 4$
Facteur1	$p = 2$
Facteur2	$q = 3$

	R1 (j=1)		R2 (j=2)		R3 (j=3)		
S1 (i=1)	43 45 46 53	$\overline{x}_{i1\,j1}$ = 46.75	41 42 43 44	$\overline{x}_{i1\,j2}$ = 42.5	42 44 46 48	$\overline{x}_{i1\,j3}$ = 45	$\overline{x}_{i1} = 44.75$
S2 (i=2)	40 40 40 43	$\overline{x}_{i2\,j1}$ = 40.75	35 37 40 40	$\overline{x}_{i2\,j2}$ = 38	37 39 40 40	$\overline{x}_{i2\,j3}$ = 39	$\overline{x}_{i2} = 39.25$
	$\overline{x}_{j1} = 43.75$		$\overline{x}_{j2} = 40.25$		$\overline{x}_{j3} = 42$		$\overline{x} = 42$

$$SCE_{totale} = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(x_{ijk} - \bar{x})^2$$

$$SCE_{totale} = (43-42)^2 + (45-40)^2 + ... + (40-40)^2 = 346$$

$$SCE_A = qn\sum_{i=1}^{p}(\bar{\bar{x}}_i - \bar{x})^2$$

$$SCE_A = 12 \times [(44.75-42)^2 + (39.25-40)^2] = 181.5$$

$$SCE_B = pn\sum_{i=1}^{p}(\bar{\bar{x}}_j - \bar{x})^2$$

$$SCE_B = 8 \times [(43.75-42)^2 + (40.25-42)^2 + (42-42)^2] = 49$$

$$SCE_{AB} = n\sum_{i=1}^{p}(\overline{x_{ij}} - \overline{x_i} - \overline{x_j} + \overline{x})^2$$

$$SCE_{AB} = 4 \times [(46.75 - 44.75 - 73.75 + 42)^2 + (42.5 - 44.75 - 40.25 + 42)^2$$

$$+ .....(39 - 39.25 - 42 + 42)^2] = 2.9$$

$$SCE_R = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(x_{ijk} - \overline{x_{ij}})^2$$

$$SCE_R = (43 - 46.75)^2 + (45 - 46.75)^2 + ...(40 - 39)^2] = 112.5$$

$$SCE_T = SCE_A + SCE_B + SCE_{AB} + SCE_R$$

$$pqn - 1 = (p-1) + (q-1) + (p-1)(q-1) + pq(n-1)$$

$$CM_T = \frac{SCE_T}{pqn - 1}$$

$$CM_A = \frac{SCE_A}{p - 1}$$

$$CM_B = \frac{SCE_B}{q - 1}$$

$$CM_{AB} = \frac{SCE_{AB}}{(p - 1)(q - 1)}$$

$$CM_R = \frac{SCE_R}{pq(n - 1)}$$

$$F_A = \frac{CM_A}{CM_R}$$

$$F_B = \frac{CM_B}{CM_R}$$

$$F_{AB} = \frac{CM_{AB}}{CM_R}$$

Source	Somme des carrés	ddl	Moyennes des carres	Fcal
Type de sol	**181.5**	**1**	**181.5**	**29**
Type d'irrigation	**49**	**2**	**24.5**	**3.91**
Interaction	**2.9**	**2**	**1.5**	**< 1**
Résidus	**112.6**	**18**	**6.26**	
Totale	**346**	**23**		

l'anova met une différence juste significative du point de vue type d'irrigation

Fcri=3.55 Fcal=3.91

l'interaction peut être considérer nulle Fab **<1**

l'a difference du point de vue type de sol est tres hautement significative

Fcri=4.41 Fcal**=29**