.

Statistiques Déductives

Bon pas de panique, il y a pas mal de calculs mais rien de bien compliqué. On apprend bien les calculs et après on s'entraine avec des exos et ça ira tout seul ! Et bon courage pour cette année !

I. Les tests d'hypothèses

a. Tout d'abord qu'est-ce que c'est ?

Ce sont des tests statistiques grâce auxquels on tente de tirer des conclusions à partir d'observations. On cherche donc à comparer deux populations.

b. Comment définir les hypothèses ?

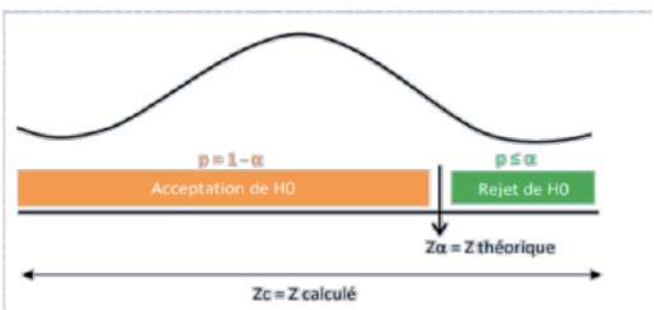
- H0 : Hypothèse nulle = aucune différence
- H1 : Hypothèse alternative = différence

Deux situations possibles :
- Situation Unilatérale : On peut seulement dire si oui ou non il y a une différence
- Situation Alternative : On peut dire qu'une des deux conditions est meilleure si différence

Le test permet par la suite de confirmer ou d'infirmer les hypothèses formulées.
En cherchant à savoir si on accepte ou rejette H0 avec un risque α .

c. Étapes de mise en œuvre d'un test d'hypothèse

1- Définir H0 et H1 *(rôles symétriques)*
2- Définir le test en fonction des données *(quantitatif ou qualitatif / grand ou petit échantillon)*
3- Choisir le risque α *(fixé à priori et le plus souvent 5%)*
4- Recueillir les données
5- Calculer Zc *(paramètre que l'on cherche)*
6- Utilisation de la règle de décision *(basé sur H0 et α et on examine Zc sur le modèle théorique = courbe de Gausse et la compare à Zt = valeur théorique lue sur la table statistique de l'écart-réduit*



Acceptation de H0	Rejet de H0
Zc < Zt	Zc > Zt
p = 1 - α	p ≤ α

7- Fixer le degré de signification de p *(= risque d'erreur réel qui se rattache à la conclusion donc à postériori = conviction avec laquelle on rejette α)*

8- Interpréter les résultats : au niveau de l'échantillon (est-ce qu'on accepte H0) et au niveau de la population (est-ce qu'on peut extrapoler)

Ces étapes sont valables pour tous les tests qui vont suivre

A savoir :

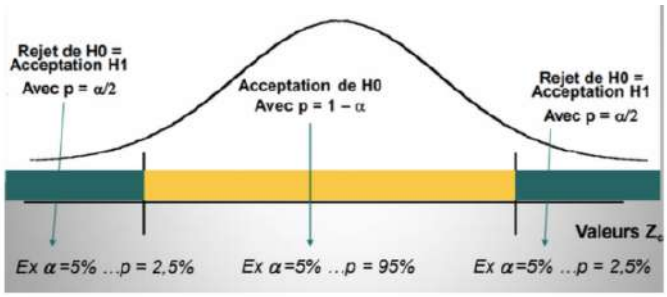
Si α = 5% alors Zt = 1,96

Si Zc < Zt : on accepte H0 et rejette H1

Si Zc > Zt : on rejette H0 et accepte H1

Si p ≤ α : on rejette H0 et accepte H1

Si H0 est accepté p = 1- α



d. Risque et puissance d'un test

α = Risque de première espèce : Probabilité de rejeter H0 alors que H0 est vraie *(rejet à tort de l'hypothèse nulle souvent fixé à 5%)*

β = Risque de deuxième espèce : Probabilité d'accepter H0 alors que H0 est fausse *(rejet à tort de l'hypothèse alternative jamais définie mais en général égale à 20%)*

1-β = Puissance du test : Probabilité de rejeter H0 alors que H0 est fausse *(probabilité de bien détecter la différence)*

On choisit toujours de maitriser α plutôt que β qui peut être pourtant important également. Rejeter H0 à tort est le plus grave.

		Décision du statisticien	
		Rejet H0	Non rejet H0
Réalité	H0 vraie	α	1- α
	H1 vraie	1-β	β

2. Liaison entre deux variables qualitatives

a. Comparaison de pourcentages

1- H0 : Pas de différence significative entre les populations pour le caractère étudié
H1 : Différence significative du caractère étudié

2- **Qualitative Binaire** *(Variable 1 : échantillon A/B - Variable 2 : Présence ou non du caractère)*

3- Souvent α = 5%

4- Pourcentages p_A et p_B de la présence du caractère dans les échantillons A et B

5- Calcul : avec $q = 1 - p$

$$\varepsilon_c = \frac{pA - pB}{\sqrt{\frac{pA.qA}{nA} + \frac{pB.qB}{nB}}}$$

6- Comparer $\mathcal{E}_c$ et $\mathcal{E}_t$ lu dans la table de l'écart-réduit

Table de l'écart réduit

	α	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	∞	2,576	2,326	2,17	2,054	1,96	1,881	1,812	1,751	1,695
0,1	1,645	1,598	1,555	1,514	1,476	1,44	1,405	1,372	1,341	1,311
0,2	1,282	1,254	1,227	1,2	1,175	1,15	1,126	1,103	1,08	1,058
0,3	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,86
0,4	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,69
0,5	0,674	0,659	0,643	0,628	0,613	0,598	0,583	0,568	0,553	0,539
0,6	0,524	0,51	0,496	0,482	0,468	0,454	0,44	0,426	0,412	0,399
0,7	0,385	0,372	0,358	0,345	0,332	0,319	0,305	0,292	0,279	0,266
0,8	0,253	0,24	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138
0,9	0,126	0,113	0,1	0,088	0,075	0,063	0,05	0,038	0,025	0,013

Pour α = 5% on a $\mathcal{E}_t$= 1,96

7- Pour α = 5%

Si **$\mathcal{E}c < \mathcal{E}t$** : on accepte H0 et rejette H1 alors **p = 1- α**

Si **$\mathcal{E}c > \mathcal{E}t$** : on rejette H0 et accepte H1 alors **p = α/2**

8- Interpréter les résultats

b. Test du X^2 *(CHI)*

1- H0 : Pas de différence dans la répartition du caractère étudié entre les populations *(pas de lien entre les deux variables étudiées)*

H1 : Différence dans la répartition du caractère étudié entre les populations *(lien entre les deux variables étudiées)*

2- Lien entre deux variables qualitatives avec n ≥ 2 *(V1 : n échantillons / V2 : n caractères étudiés)*

3- Souvent α = 5%

4- Effectifs Observés

V2\V1	Modalité A	Modalité B	Totaux
Modalité 1	O_1	O_2	O_1+O_3
Modalité 2	O_3	O_4	O_2+O_4
Totaux	O_1+O_3	O_2+O_4	O_1+O_2 +O_3+O_4

5- Calculer X^2

o_i = *données observées*

c_i : *données calculées* = $\frac{(ligne\ i)\times(colonne\ i)}{Effectif\ total}$

$$X_c^2 = \frac{\sum(o_i - c_i)}{c_i}$$

6- Comparaison de X_c^2 à X_t^2 lue dans la table du X^2 fonction du risque et du degré de liberté (ddl) = *nombre minimal de valeur dans une série qui permet de calculer les valeurs manquantes lorsqu'on dispose des totaux*

DDL = (nombre de lignes - 1) x (nombre de colonnes -1)

7- Comparer X²c et X²t

ddl \ α	0,9	0,5	0,3	0,2	0,1	0,05	0,02	0,01	0,001
1	0,016	0,455	1,074	1,642	2,706	**3,841**	5,412	6,635	10,827
2	0,211	1,386	2,408	3,219	4,605	5,991	7,824	9,21	13,815
3	0,584	2,366	3,665	4,642	6,251	7,815	9,837	11,345	16,266
4	1,064	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,467
5	1,61	4,351	6,064	7,289	9,236	11,07	13,388	15,086	20,515
6	2,204	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457
7	2,833	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322

Pour α = 5% et ddl =1 on a $X_t^2 = 3{,}841$

8- Interpréter les résultats

3. Liaison entre deux variables : qualitative et quantitative

a. Comparaison de moyennes pour séries indépendantes

1- H0 : Les moyennes ne sont pas différentes
H1 : Les moyennes sont différentes
2- Lien d'une variable qualitative avec une variable quantitative indépendante avec n_A et $n_B \geq 30$ = *grands échantillons*
3- Souvent 5%
4- Calcul des moyennes m_A et m_B et des écarts-types s_A et s_B
5- Calcul de ε

$$\varepsilon = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

6- Voir Table de l'écart réduit
7- Comparer ε_α et ε
8- Interpréter les résultats

b. Test t de Student pour séries indépendantes

1- H0 : Les moyennes ne sont pas différentes
H1 : Les moyennes sont différentes
2- Lien d'une variable qualitative avec une variable quantitative indépendante
avec n_A et $n_B < 30$ = *petits échantillons*
3- Souvent 5%
4- Calcul des moyennes m_A et m_B et des écart-types s_A et s_B
5- Calcul de t

$$t = \frac{m_1 - m_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

6- Lecture dans la table t de Student
Avec **DDL = (n1-1) + (n2-1)**

d.d.l. \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,703	1,100	1383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437

7- Comparer t et t_α
8- Interpréter les résultats

c. Comparaison de moyennes pour séries appariées = méthode des couples

1- H0 : Les moyennes ne sont pas différentes
H1 : Les moyennes sont différentes

2- Lien d'une variable qualitative avec une variable *quantitative (1 échantillon mais 2 valeurs par personne)* avec n_A et $n_B \geq 30$ = *grands échantillons*
3- Souvent 5%
4- Calcul de la moyenne m_d et de l'écart-type s_d avec d la différence de résultat pour un même sujet
5- Calcul de ε

$$\varepsilon = \frac{m_d}{\sqrt{\frac{s^2}{n}}}$$

6- Voir Table de l'écart réduit
7- Comparer ε_α et ε
8- Interpréter les résultats

d. Test t de Student pour séries appariées = méthode des couples

1- H0 : Les moyennes ne sont pas différentes
H1 : Les moyennes sont différentes
2- Lien d'une variable qualitative avec une variable *quantitative (1 échantillon mais 2 valeurs par personne)* avec n_A et $n_B < 30$ = *petits échantillons*
3- Souvent 5%
4- Calcul de la moyenne m_d avec d la différence de résultat pour un même sujet et la variance de la différence s^2
5- Calcul de t

$$t = \frac{m_d}{\sqrt{\frac{s^2}{n}}}$$

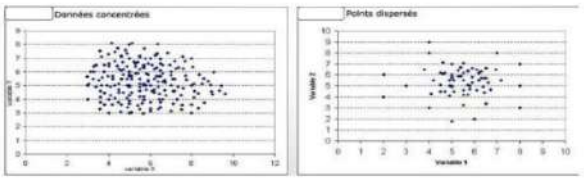
6- Voir table t de Student avec **ddl = n – 1**
7- Comparer t et t_α
8- Interpréter les résultats

4. Liaison entre deux variables quantitatives

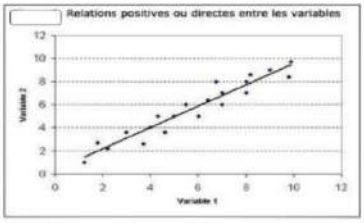
a. Corrélation et régression

Corrélation : évaluation de la liaison entre deux variables quantitatives
Régression : méthode qui permet l'explication mathématique des relations entre variables observées

Nuage de points :



Droite de régression : permet de visualiser si une des 2 variables est **dépendante** de l'autre



V1 en x
V2 en y
Soit la droite y=f(x)

Droite de régression = Droite des moindres carrées *: passe au plus près de chaque point du nuage -> elle permet de prédire la valeur d'une variable en connaissant l'autre.*

b. Test du coefficient de corrélation

1- H0 : Les variables sont indépendantes entre elles *(Pas de lien)*
   H1 : Les variables sont dépendantes = corrélées *(lien entre les deux variables étudiées)*
2- Variables quantitatives avec n ≥ 10 par échantillon
3- Souvent α = 5%
4- Recueil de x_i et y_i pour chaque valeur pour chaque échantillon
   et de leur moyenne m_x et m_y
5- Calcul de r = Coefficient de corrélation = pente de la droite *(TJS < 1)*
   Si r = 0 alors pas calculable *(non lié)*
   Si r > 0 : Les variables varient dans le même sens *(liaison positive)*
   Si r < 0 : Les variables varient en sens opposé *(liaison négative)*

$$r = \frac{\sum(x_i - m_x)(y_i - m_y)}{\sqrt{\sum(x_i - m_x)^2 \sum(y_i - m_y)^2}} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{(\sum x^2 - \frac{(\sum x)^2}{n})(\sum y^2 - \frac{(\sum y)^2}{n})}}$$

6- Lecture de la table du coefficient de corrélation avec **DDL = n -2**

d.d.l. \ α	0,10	0,05	0,02	0,01
1	0,9877	0,9969	0,9995	0,9999
2	0,9000	0,9500	0,9800	0,9900
3	0,8054	0,8783	0,9343	0,9587
4	0,7293	0,8114	0,8822	0,9172
5	0,6694	0,7545	0,8329	0,8745
6	0,6215	0,7067	0,7887	0,8343
7	0,5822	0,6664	0,7498	0,7977
8	0,5494	0,6319	0,7155	0,7646
9	0,5214	0,6021	0,6851	0,7348
10	0,4973	0,5760	0,6581	0,7079
11	0,4762	0,5529	0,6339	0,6835
12	0,4575	0,5324	0,6120	0,6614
13	0,4409	0,5139	0,5923	0,6411
14	0,4259	0,4973	0,5742	0,6226

7- Comparer $|r|$ et r_α
8- Extrapolation possible uniquement si tirage au sort !
Corrélation ≠ Causalité

5. Tests non paramétriques

a. Définition

Utilisé en cas d'effectif très faible soit **4 < n < 12** avec un caractère quantitatif c'est le seul test possible. Test avec forte robustesse *(= utile alors que la population ne se distribue pas normalement*) car transforme les données quantitatives en mesures ordinaires *(=rangs)*

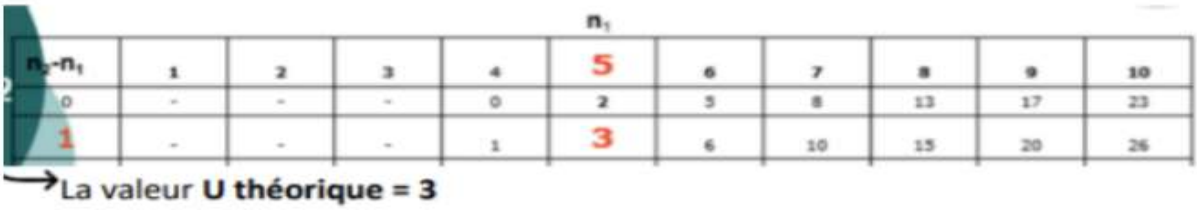
b. Test U de Mann et Whitney

1- H0 : Les moyennes sont proches *(Pas de différence)*
H1 : Les moyennes ne sont pas proches *(différence)*
2- **Une variable quantitative et une variable qualitative** avec n_A et ou n_B < 12
3- Souvent α = 5%
4- Recueil des valeurs de A et B puis rangement croissant
Attention les valeurs égales comptent 1/2

Nbr	1	5	7	10	13	23	34	56	67	134	235
Grp	A	A	B	A	B	A	A	B	A	B	B
U	0	0	2	1	3	2	2	5	3	6	6

5- Calcul de U_{AB} et U_{BA}
U_{AB} = 2+3+5+6+6 = 22
U_{BA} = 0+0+1+2+2+3 = 8
RQ: $U_{AB} + U_{BA} = n_A + n_B = 22+8 = 6x5 = 30$
U_c est la plus petite somme entre U_{AB} et U_{BA} de même pour n_1

6- *Avec $n_1 = 5$ et $n_2 - n_1 = 1$ alors Ut = 3*



n_2-n_1	n_1 = 1	2	3	4	**5**	6	7	8	9	10
0	-	-	-	0	2	5	8	13	17	23
1	-	-	-	1	**3**	6	10	15	20	26

La valeur **U théorique = 3**

7- Comparer U et $U_{théo}$

8- Interpréter les résultats

c. Test du coefficient r' de Spearman

1- H0 : Pas de lien entre les variables
H1 : Lien entre les variables

2- Deux **variables quantitatives** avec n < 12

3- Généralement 5%

4- Recueil des valeurs, classement par ordre croissant, Calcul de la différence de rang attribué

5- Calcul de r' *(tjs < 1)*
di = différence de rang entre x et y pour un couple i
Si r > 0 : Les variables varient dans le même sens *(liaison positive)*
Si r < 0 : Les variables varient en sens opposé *(liaison négative)*

$$r' = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

6- On cherche r' fonction de α et n

n \ α	0.2	0.1	0.05	0.02	0.01	0.002
4	1.000	1.000	–	–	–	–
5	0.800	0.900	1.000	1.000	–	–
6	0.657	0.829	0.886	0.943	1.000	–
7	0.571	0.714	0.786	0.893	0.929	1.000
8	0.524	0.643	0.738	0.833	0.881	0.952
9	0.483	0.600	0.700	0.783	0.833	0.917
10	0.455	0.564	0.648	0.745	0.794	0.879
11	0.427	0.536	0.618	0.709	0.755	0.845
12	0.406	0.503	0.587	0.678	0.727	0.818

7- Comparer $|r'|$ et r'_α

8- Si H1 vrai alors les deux séries sont corrélées

6. Synthèse

	Test	
	Paramétriques	Non paramétrique
Comparaison de 2 échantillons indépendants	♠ **Test t de Student** ♠ **Test de comparaison de moyennes**	♠ **Test de Mann-Whitney**
Comparaison de 2 échantillons appariés	♠ **Test t de Student pour séries appariées** ♠ **Test de comparaison de moyennes pour séries appariées**	♠ **Test de Wilcoxon**
Test de corrélation	♠ **Test du coefficient r**	♠ **Test du coefficient r' de Spearman**

Effectif	Données quantitatives	Données qualitatives	Données qualitatives et quantitatives
n ≥ 30	• **Coefficient de corrélation r** • r' de Spearman	• Comparaison de pourcentages • Chi-2	• **Comparaison de moyennes** • T de Student • U Mann et Whitney
30 > n ≥ 12	• **Coefficient de corrélation r** • r' de Spearman	• Comparaison de pourcentages • Chi-2	• **T de Student** • U Mann et Whitney
12 > n > 4	• **r' de Spearman**	• Comparaison de pourcentages • Chi-2	• U Mann et Whitney

Petite dédicace rapide car il est 1h du matin, que j'ai sommeil et que j'ai déjà dépassé la dead line mais bon je voulais vous faire un joli cours.
Dédicace au tuteur de l'an dernier sur qui j'ai pris exemple et à qui j'ai piqué quelques phrases
Dédicace à ma maman parce que bon je serais pas là sans elle quand même
Dédicace à nos chef Tut
Dédicace à toute la famille Biostat
Dédicace à vous les P1 à qui je souhaite beaucoup de courage et en qui je crois pour cette année