STATISTIQUES DÉDUCTIVES

I. Généralités sur les tests d'hypothèse

Le but principal des statistiques déductives est de tirer des conclusions à partir des observations. Le plus souvent, on essaiera de comparer 2 groupes pour un caractère donné

Ex : on peut comparer les notes à l'épreuve de biostat de l'année dernière à celles de cette année. On se pose la question : y a-t-il une différence entre ces deux groupes (ou deux années ici)

A. Définition des hypothèses

En statistiques descriptives on travaille à partir de 2 hypothèses :

H0	H1
⇒ L'hypothèses **nulle** ⇒ Il n'y a **pas** de différence entre les 2 groupes ⇒ Les fluctuations observées sont dues au **hasard**	⇒ L'hypothèse **alternative** ⇒ Il y a une **différence significative** entre les 2 groupes (ou le groupe 1 est meilleur que le 2 ou …) ⇒ Les fluctuations observées ne sont pas dues au hasard

Un test, c'est une technique qui permet de décider si on accepte ou rejette H0 en ayant fixé le risque d'erreur α accompagnant cette décision

B. Étapes d'un test hypothèse +++

1. Définir **H0** et **H1**
2. Choisir le **test** en fonction du type de données (qualitative, quantitative, nombre de données)
3. Fixer le **risque** α (souvent 5%)
4. Recueillir les données
5. Calculer Z
6. Utiliser la règle de **rejet/acceptation de H0** : comparer le Z_c (Z calculé) au Z_t (Z théorique) dont on connait la distribution
7. Fixer le **risque d'erreur réel** (à posteriori)
8. **Interpréter** les résultats : interprétation **statistique + médicale**

C. Risque

Le risque de 1ere espèce ou **risque** α : c'est la probabilité de rejeter H0 si H0 est vraie (à tort).
Ce risque est maitrisé, on le fixe à l'avance (en général à 5%)

Le risque de 2nde espèce ou **risque ß** : c'est la probabilité d'accepter H0 si H0 est fausse.
Ce risque est négligé, il peut être très élevé (en général ß = 20%)

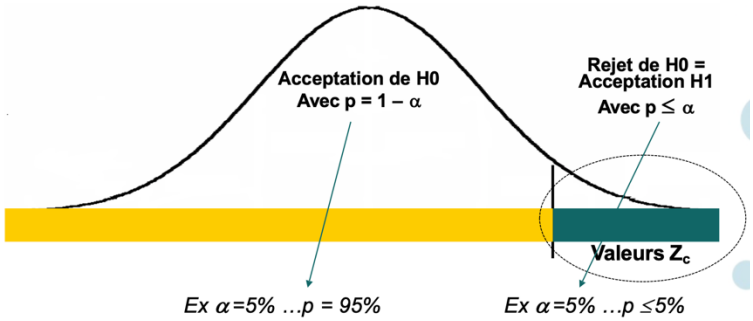
La **puissance** du test, 1 - ß : on rejette H0 avec H1 vraie

La règle de rejet du test est définie seulement à partir de α et de H0. Entre 2 alternatives, on choisira pour H0 l'hypothèse qu'il serait le plus grave de rejeter à tort

	Rejet H0	Non rejet H0
H0 vraie	α	1-α
H1 vraie	1-ß	ß

D. Interprétation graphique

Le paramètre Z suit une distribution en forme de Gauss



Pour arriver à une conclusion on doit :

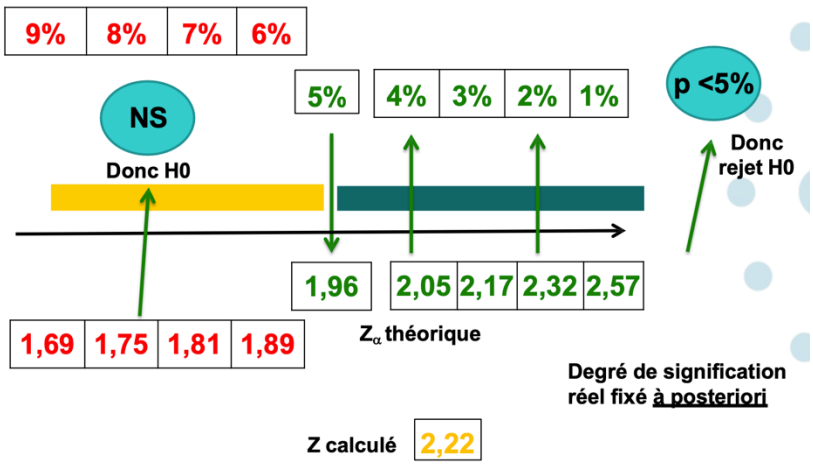
1. Fixer le risque α **à priori**
2. Chercher le Z_t dans la table *(vu plus loin)*
3. Calculer Z_c grâce aux formules
4. Comparer Z_c et Z_t

Ici on peut arriver à 2 situations :

$Z_c < Z_t$	$Z_c > Z_t$
Acceptation de H0	Rejet de H0
p = 1 - α	p ≤ α

5. Fixer le degré de signification p **à posteriori**

Le statisticien fixe le risque α à priori mais dans certains cas il est possible d'avoir une précision d'étude supérieur à celle fixée au départ.



1. α = 5%
2. Zα = 1,96
3. Z_c = 2,22
4. 2,22 > 1,96 donc on rejette H0
5. Pour α = 1% Zα = 2,57 or 2,22<2,57 donc on ne rejette pas H0 à 1%. La précision n'a pas augmenté.

p<5%

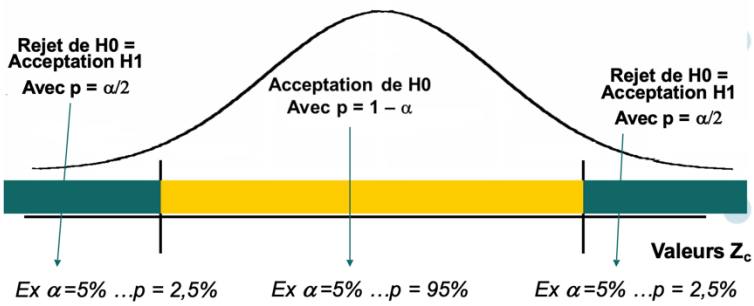
Ici on pourrait dire qu'on rejette H0 à 3% mais en pratique on utilise seulement 1 et 5%

Si je rejette ou accepte H0 à tous les seuils, le test n'est pas très discriminant ou non significatif

On peut se retrouver face à 2 situations :

⇒ Situation **unilatérale** : le rejet de H0 permet seulement de dire qu'il y a une différence significative entre les 2 situations. C'est la situation la plus fréquente.

⇒ Situation **bilatérale** : L'acceptation de H1 permet de déterminer laquelle des situations est la meilleure



Ex : Si on compare 2 traitements A et B, en situation unilatérale, en rejetant H0 on pourra dire qu'il y a une différence significative entre les 2 traitements. En situation bilatérale, on pourra dire qu'il y a une différence significative et que le traitement A est meilleur que le B

E. Big data

→ Et si les données étaient le pétrole du 21ème siècle ?
Nous générons et détenons quantités d'info personnelles : alimentation, achats, contributions réseaux sociaux, goûts, préférences, recherches sur Google, santé connectée, ...
Ces données sont éparses mais captées par différents intervenants sur Internet.

Dans le **domaine de la santé**, des études épidémiologiques diverses sont lancées (pour le meilleur et pour le pire ?) : aux USA, des société privées analysent ces data et en tirent des conclusions. Par exemple, ils proposent à des femmes l'ablation des 2 seins car leur profil génétique comparé à celui de milliers d'autres femmes suppose un risque accru de cancer du sein.

Les **objets connectés** (bracelets, balances, tee-shirts, fauteuils, iwatch ,..) permettent de suivre sa propre forme physique, la comparer à ce qu'elle devrait être (mais qui définit les normes).
Mais alimentent aussi de manière continue ces fameuses Big Data

L'utilisation de ces masses de données remet en cause certaines théories statistiques et la notion d'échantillonnage.

Jusqu'à aujourd'hui les données recueillies dans les **études cliniques** sont des données **démographiques** (sexe, âge), **cliniques** (poids, taille, diagnostique, traitement, dose, durée), **biologiques**, ... Jamais de données de type **psychologique** ou **émotionnel**, …

Les Big Data permettent de recouper et analyser TOUS ces types de données et de remettre en cause certaines conclusions ou décisions.
De plus, un échantillon traditionnel est un effectif de quelques dizaines, au mieux quelques centaines d'individus, représentant des populations cibles souvent de plusieurs centaines de milliers d'individus.
Grâce aux Big Data, l'effectif de l'échantillon observé et étudié est de l'ordre de la population cible
Cela règle le problème du nombre de sujets à étudier.

II. Lien entre 2 variables qualitatives

À partir d'ici les formules ne sont pas à connaître sauf les formules « simples » comme le chi 2

On se demande si le pourcentage d'individu possédant un caractère x dans un groupe A est le même que le pourcentage d'individu possédant le caractère x dans le groupe B. Le caractère x est ici qualitatif (couleur des yeux, porteur de lunettes, …)

A. Test de comparaison des pourcentages (Tout effectif)

Ici le paramètre Z est l'écart réduit ε

⇒ ε_t vient de la table de l'écart réduit

⇒ $\varepsilon_c = \frac{p_A - p_B}{\sqrt{\frac{p_A q_A}{n_A} + \frac{p_B q_B}{n_B}}}$ avec $q_A = 1 - p_A$

Si $\varepsilon_c > \varepsilon_t$ -> rejet de H0

Comment lire ε_t dans la table ?

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
0,5ε	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

Table pour les petites valeurs de la probabilité

0,001	0,000 1	0,000 01	0,000 001	0,000 000 1	0,000 000 01	0,000 000 001
3,2905	3,8905	4,41717	4,89164	5,32672	5,73073	6,10941

On cherche ε_t en fonction d'α

On regarde les dixièmes d'α sur les lignes et les centièmes sur les colonnes. Le ε_t sera à l'intersection

Ex : Pour $\alpha = 5\% = 0,05$: on regarde 0,00 pour les lignes et 0,05 pour les colonnes : $\varepsilon_t = 1,96$

Pour $\alpha = 0,1\% = 0,001$: on regarde la table pour les petites valeurs $\varepsilon_t = 3,29$

Exemple :

Soit 2 groupes de 200 enfants. Crèche : 200 enfants, 130 rhinos. Maison : 200 enfants, 96 rhinos

Le mode de garde influe-t-il sur le risque de rhinopharyngites ?

1. *H0 : il n'y a pas de différence entre les 2 modes de garde vis-à-vis du développement de rhinos. H1 : il y a une différence*
2. *Caractère 1 : gardé en crèche ou à domicile : qualitatif*
   Caractère 2 : développer une rhinopharyngite ou non : qualitatif
   → test de comparaison des pourcentages
3. *$\alpha = 5\%$*
4. *Recueil des données*
5. *$p_A = 65\%$ $p_B = 48\%$. $\varepsilon_c = 3,4$*
6. *3,4 > 1,96 : on rejette H0 au seuil 5%*
7. *3,4 > 3,3 donc on rejette H0 au seuil 0,001*
8. *Sur cet échantillon, le risque de rhino est supérieur chez les enfants gardés en crèche. On ne peut pas généraliser car il n'y a pas eu de tirage au sort et il manque des infos sur les enfants (précision du mode de garde à domicile, du revenu des parents, …)*

<u>Remarque</u> : On travaille sur des proportions donc il n'est pas nécessaire d'avoir le même effectif dans le groupe A que dans le groupe B

B. Test du X^2 (Tout effectif)

On utilise de préférence ce test si notre tableau de données à plus de 2 lignes (ou 2 colonnes)
Ici le paramètre Z est X^2

⇒ X^2_t vient de la table du X^2

⇒ **$X^2_c = \sum \frac{(o_i - c_i)^2}{c_i}$** Avec o_i les données observées et c_i les données calculées

Si $X^2_c > X^2_t$ -> Rejet de H0

DDL = (nombre de lignes – 1) * (nombre de colonnes – 1)

Comment lire X^2_t dans la table ?

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
8	3,49	7,344	9,524	11,03	13,362	15,507	18,168	20,09	26,125
9	4,168	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	4,865	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588
11	5,578	10,341	12,899	14,631	17,275	19,675	22,618	24,725	31,264
12	6,304	11,34	14,011	15,812	18,549	21,026	24,054	26,217	32,909
13	7,042	12,34	15,119	16,985	19,812	22,362	25,472	27,688	34,528
14	7,79	13,339	16,222	18,151	21,064	23,685	26,873	29,141	36,123
15	8,547	14,339	17,322	19,311	22,307	24,996	28,259	30,578	37,697
16	9,312	15,338	18,418	20,465	23,542	26,296	29,633	32	39,252
17	10,085	16,338	19,511	21,615	24,769	27,587	30,995	33,409	40,79
...									

X^2_t dépend d'α et du DDL
Le DDL ou **degré de liberté** est le nombre minimal de valeur nécessaire dans une série pour pouvoir calculer toutes les autres
On cherche le ddl sur les lignes et α sur les colonnes
Ex : Si α = 5% et DDL = 1 alors X^2_t = 3,8

<u>Exemple</u> : On cherche à savoir si l'exposition professionnelle au benzène peut entrainer une leucémie

	Leucémie	*Non leucémie*	***Total***
Expo	*15*	*485*	*500*
Non expo	*20*	*980*	*1000*
Total	*35*	*1465*	*1500*

1. *H0 : il n'existe pas de lien entre l'exposition au benzène et les leucémies*
2. *Variable 1 : leucémie ou non : qualitatif*
   Variable 2 : Exposé ou non : qualitatif
   → test du X^2
3. *α = 5%*

Les valeurs 15, 20, 485 et 980 sont des valeurs observées. On va maintenant chercher les valeurs calculées par un modèle théorique :
Il y a 35 malades pour 1500 personnes au total soit 2,33% de malade. On va appliquer ce pourcentage aux exposés et aux non exposés.
2,33% de 500 (les exposés) = 11,65 malades chez les expos (chiffre théorique)
2,33% de 1000 (les non-expos) = 23,35
Il y a 1465 non malades pour 1500 personnes au total soit 97,67%. On applique ce pourcentage aux expos et aux non-expos :
97,67% de 500 = 488,3
97,67% de 1500 = 976,7
$X^2_c = \frac{(15-11,65)^2}{11,65} + \frac{(20-23,35)^2}{23,35} + \frac{(485-488,3)^2}{488,3} + \frac{(980-976,7)^2}{976,7} = 1,42$
X^2_t *: ddl = (2-1) * (2-1) = 1 donc* $X^2_t = 3,84$
$X^2_c < X^2_t$ *donc on accepte H0 au seuil 0,05*
Il n'existe pas de relation entre l'exposition au benzène et les leucémies

III. Lien entre variables qualitatives et quantitatives

On se demande si en moyenne la taille des individus d'une population A coïncide avec la taille des individus d'une population B

A. Test de comparaison de moyennes (n1 et n2 > 30 : grands échantillons)

Ici le paramètre Z est l'écart-réduit ε

⇒ ε_t vient de la table de l'écart-réduit

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

Si ε_c -> ε_t -> Rejet de H0

<u>Exemple</u> : On cherche à comparer le taux de T3 libre chez les femmes prenant un contraceptif oral et celles qui n'en prennent pas. Après tirage au sort on obtient :
Femmes sans c.o : $n_1 = 50$, $m_1 = 2$ nmol et $s_1 = 0,35$ nmol
Femmes avec c.o : $n_2 = 33$, $m_2 = 2,5$ nmol et $s_2 = 0,3$ nmol

1. *H0 : les moyennes ne sont pas différentes, ce sont 2 estimateurs du taux de T3 libre chez la femme en général*
2. *Variable 1 : prise ou non de la pilule : qualitatif*
   Variable 2 : dosage de T3 : quantitatif
   n_1 *et* $n_2 > 30$
   → *test de comparaison de moyennes*
3. $\alpha = 5\%$
4. $\varepsilon_t = 1,96$
5. $\varepsilon_c = 6,94$
6. $\varepsilon_c > \varepsilon_t$ *donc rejet de H0*
7. $p < 0,0001$
8. *Il y a eu TAS donc le résultat est généralisable : la prise de c.o augmente le taux de T3 libre*

B. Test T de student (n1 ou n2 < 30 : petits échantillons)

Ici le paramètre Z est t

⇒ t_t est lu dans la table du t de student

⇒ $t_c = \frac{m_1 - m_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$: c'est presque la même formule que pour la comparaison de moyenne mais on utilise seulement l'écart-type s car il est moins significatif ici

⇒ L'écart-type s = $\sqrt{\frac{\sum(x_i - m_1)^2 + \sum(x_j - m_2)^2}{(n_1 - 1) + (n_2 - 1)}}$ *(Trop compliqué à calculer on vous le donnera dans l'énoncé)*

Si $t_c > t_t$ -> rejet de H0

DDL = $(n_1 - 1) + (n_2 - 1)$

Précision sur le ddl :

2	3	5	12	10		7	8	**51**
2	3	5	12	10			8	**51**

Avec n-1 valeur et le total, on peut trouver qu'il manque 4. Avec n-2 valeurs, on ne peut pas trouvez les deux valeurs manquantes. Le degré de liberté est donc de n-1 ici.

Exemple : Soit 15 femmes obèses et 12 femmes de poids normal. On mesure le taux de corticoïde sanguin moyen dans chaque groupe. L'obésité a-t-elle une influence sur le taux de corticoïde ?
$n_1 = 15$ $m_1 = 6,3$ $s_1 = 1,8$
$n_2 = 12$ $m_2 = 4,5$ $s_2 = 1,6$

1. *H0 : m_1 et m_2 ne sont pas différents dans les 2 groupes*
2. *Variable 1 : obèse ou non : qualitatif*
   Variable 2 : taux de corticoïde : quantitatif
   n_1 ou $n_2 < 30$ → T de student
3. *$\alpha = 5\%$*
4. *DDL = 15 + 12 – 2 = 25 donc $T_t = 2,06$*
5. *$T_c = 2,92$*
6. *$T_c > T_t$ donc on rejette H0 au seuil 5%*
7. *p < 1% après lecture dans la table. On rejette H0 à 1% à posteriori*
8. *Il existe une relation claire entre l'obésité et le taux de corticoïde au niveau de cet échantillon*

C. Séries appariées ou méthode des couples

On utilise cette méthode lorsque les 2 échantillons étudiés ne sont pas indépendants

Série indépendante : les 2 groupes comparés sont distincts et indépendants (sans lien)
Ex : Par TAS on prend un groupe 1 à qui on fait une prise de sang puis un groupe 2 à qui on fait aussi une prise de sang. Il n'y a pas de lien entre le groupe 1 et le groupe 2
Série appariée : les 2 groupes comparés ne sont pas distincts et indépendants (liés)
Ex : On fait une prise de sang à un groupe puis une prise de sang à ce même groupe 6 mois plus tard. Il y a un lien entre les premiers et les derniers résultats car l'analyse sanguine et propre à chacun

Si n > 30 on utilise le test de comparaison des moyennes : $\varepsilon = \frac{m_d}{\sqrt{\frac{s^2}{n}}}$ *Avec d différence de résultat pour un même sujet,*

Si n < 30 on utilise le test t de student : $T = \frac{m_d}{\sqrt{\frac{s^2}{n}}}$ *m_d moyenne des d, s variance des d, n nombre de couples*

Le reste de la méthodologie est identique

Exemple : On souhaite évaluer l'effet d'une substance S capable de désintoxiquer les fumeurs. On considère par TAS 2 groupes de 40 fumeurs. L'un reçoit la substance S, l'autre reçoit le placebo P. Le traitement dure 2 mois. La consommation de cigarette par jour (C) est notée avant et après traitement.

	S (n=40)		P (n = 40)	
	m_1	s_1^2	m_2	s_2^2
C avant ttt	19,5	54,2	16,5	35,6
C après ttt	5,4	30,4	3,8	20,1
Variation de C	14,1	9,1	12,7	8,9

1. Quelle est la 1ère précaution à prendre ?

Les 2 groupes doivent être comparables pour les paramètres qui peuvent influencer le ttt : âge, sexe, CSP (catégorie socio-professionnelle), consommation par jour

On compare donc les consommations moyennes avant tt dans les 2 groupes

- *H0 : les moyennes de consommation sont équivalentes dans les 2 groupes*
- *Variable 1 : S ou P = qualitative, Variable 2 : C = quantitative*
- *Échantillons indépendants → test de comparaison des moyennes*
- *$\varepsilon_c = 2 > 1,96$*
- *On rejette H0 avec un risque $\alpha = 5\%$.*

Il y a donc une différence significative de la consommation moyenne de cigarette par jour dans les 2 groupes. On fume plus dans le groupe S (situation bilatérale). Il faut en tenir compte lors de l'étude de la variation de consommation avant et après ttt

2. Dans le groupe placebo, la consommation moyenne diffère-t-elle avant et après ttt ?

- *Variable 1 : avant après ttt = qualitatif, Variable 2 : C = quantitative.*
- *Échantillon non indépendants → méthode des couples ; n > 30 → test de comparaison des moyennes*
- *$\varepsilon_c = 26,9 > 1,96$: rejet de H0*

Il y a une différence très significative (p < 0,001) entre C avant et après ttt dans le groupe placebo. Il y a un effet psychologique : l'envie de profiter de l'étude pour arrêter de fumer

3. Les 2 groupes diffèrent-ils dans leurs conso moyenne après ttt ?

- *H0 : les moyennes de consommation sont les mêmes dans les 2 groupes*
- *Variable 1 : S ou P = qualitative, Variable 2 : C = quantitative*
- *Échantillons indépendants et n > 30 → test de comparaison des moyennes*
- *$\varepsilon_c = 1,42 < 1,96$: on accepte H0 au seuil 5%*

Il n'existe pas de différence significative entre les 2 groupes pour la consommation après ttt

4. Les 2 groupes diffèrent-ils pour la variation de consommation avant et après ttt ?

Il faut comparer les variations dans les 2 groupes pour prouver l'efficacité de la substance S

- *H0 : il n'existe pas de différence entre les variations de consommation dans les 2 groupes*
- *Variable 1 : S ou P = qualitative, Variable 2 : C = quantitative*
- *n > 30 → test de comparaison des moyennes*
- *$\varepsilon_c = 2,09 > 1,96$*
- *Rejet de H0 au risque 5%*

Il existe une différence significative entre les variations de consommation dans les 2 groupes (p<5%)
Conclusion : Il y a eu TAS donc le résultat est généralisable

***Conclusion générale** : Il n'y a pas de différence de consommation après traitement (Q3) mais il y avait une différence avant traitement (le groupe S fumait plus : Q1). On peut donc cire qu'il y a une efficacité du traitement S pour désintoxiquer les fumeurs*

IV. Lien entre 2 variables quantitatives

A. Corrélation et régression

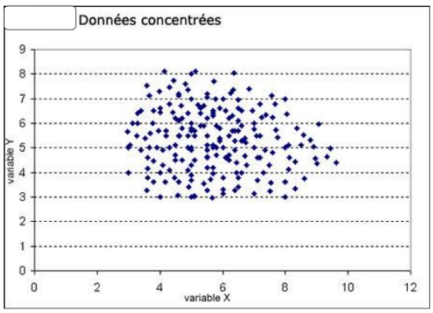
Corrélation : évaluation de la liaison entre 2 variables quantitatives
Régression : méthode mathématique qui permet d'expliquer les relations entre les variables observées

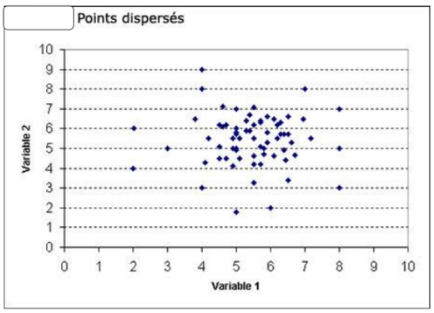
B. Représentation des données

En variable x, on met la variable explicative. En variable y, on met la variable à expliquer.

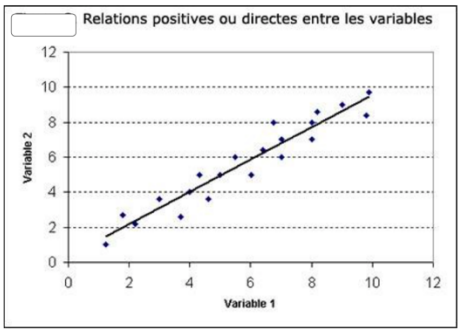
Nuage de point :

Il n'y a pas de relation entre x et y





Droite de régression : elle permet de visualiser si l'une des 2 variables est dépendante de l'autre.
La droite de régression est aussi appelée droite des moindres carrées car elle passe au plus près de chaque point du graphe
Dans ce cours on ne parle que de régression linéaire car on a choisi d'avoir une droite et pas un polynôme



C. Étude de la liaison entre caractères quantitatifs

Ex : la capacité respiratoire des enfants est-elle dépendante de la consommation de cigarettes de leurs mères ?
Le poids des bébés à la naissance est-il lié à l'âge de la mère ?

Une droite de régression peut permettre de prédire certaine valeur de y à partir d'une valeur x
Plus on a de valeurs, plus notre droite permettra de prédire les valeurs suivantes de manière précise.
Avec seulement 3 valeurs, la 4eme valeur sera prédite de manière imprécise

Exemple : On a un échantillon de 10 sujets. On recueil chez chacun leur âge et leur concentration de cholestérol. Est-ce que le taux de cholestérol est lié à l'âge ?

X âge	*30*	*60*	*40*	*20*	*50*	*30*	*40*	*20*	*70*	*60*
Y chol	*1,6*	*2,5*	*2,2*	*1,4*	*2,7*	*1,8*	*2,1*	*1,5*	*2,8*	*2,6*

1. *H0 : le taux de cholestérol n'est pas lié à l'âge*
2. *Variable 1 : Age = quantitatif*
   Variable 2 : taux de cholestérol = quantitatif
   → Test du coefficient de corrélation
3. *α = 1%, DDL = 10-2 = 8 donc r_t = 0,76*
4. *r_c = 0,955 > r_t*
5. *On rejette H0 au seuil 1%*

On obtient une relation significative au seuil 1% : plus l'âge augmente, plus le taux de cholestérol augmente. Le résultat n'est pas généralisable car on a seulement 10 individus sans TAS

Corrélation ≠ causalité : Si d'un point de vue mathématique on a obtenu une corrélation entre des paramètres statistiques, cela n'implique pas une relation de cause à effet entre les paramètres.
Corrélation : il existe un lien : *l'âge et le cholestérol sont liés*
Causalité : l'un est la conséquence de l'autre : *l'âge cause le cholestérol*

V. Tests non paramétriques

Test paramétrique : test à forte contrainte car il n'est fiable que si les données suivent une distribution selon une loi normale
Test non paramétrique : test qui ne précise pas les conditions que doivent remplir les paramètres de la population dont a été extrait l'échantillon

On utilise obligatoirement un test non paramétrique quand les effectifs sont très faibles (4<n<12)

Pour les variables quantitatives, on utilise obligatoirement un test non paramétrique si les effectifs sont inférieurs à 5 car les populations ne sont plus distribuées normalement.

A. U de Mann et Whitney

Le tets U de Mann et Whitney ou Wilcoxon-Mann-Whitney ou test de la somme de rangs de Wilcoxon permet de tester l'hypothèse selon laquelle les moyennes des 2 groupes de données sont proches.

On a 2 échantillons E_1 et E_2 de taille n_1 et n_2 indépendants

1. On réunit les valeurs des 2 échantillons
2. On trie la réunion en ordre **croissant**
3. Pour chaque valeur issue de E_1, on compte le nombre de valeur de E_2 situées **après** (s'il y a des valeurs égales, elles ne valent que $\frac{1}{2}$) *(peu d'importance entre avant ou après tant qu'on fait la même chose tout le long)*
4. La somme de ces nombres vaudra u_1
5. On échange les rôles des 2 échantillons pour trouver la somme u_2
6. Le u de Mann et Whitney est le **minimum** entre u_1 et u_2
7. On compare u avec u_t de la table

On note U la variable aléatoire associée (pour pouvoir parler de probabilité on doit parler d'une variable aléatoire)
On lit dans la table le nombre $m\alpha$ tel que $P(U \leq m\alpha) = \alpha$
On **rejette H0** au risque α si **$u \leq m\alpha$** sinon on accepte H0

Si $U_c > U_t$ -> on accepte H0

Si les effectifs sont grands (n_1 et $n_2 > 20$ en général), U suit approximativement la loi normale

Exemple : On répartit par tirage au sort 20 malades dépressifs en 2 groupes de 10. Le 1er groupe reçoit la molécule et le 2ème reçoit le placebo. On évalue les patients sur une échelle de 0 à 50 (pas déprimé -> très déprimé). Les patients sont évalués avant puis après ttt (J28). La nouvelle molécule a-t-elle un effet anti-dépresseur ?

Témoins	J0	34	30	25	27	31	24	28	30	35	26
	J28	31	28	26	25	24	25	26	27	32	25
Traités	J0	27	32	30	28	25	33	29	31	32	29
	J28	22	25	23	26	20	27	21	26	25	23

1. Y a-t-il un effet placebo ?

- *H0 : le placebo n'a aucun effet, les scores J0 ne diffèrent pas des scores J28*
- *Variable 1 : J0 – J28 -> qualitatif, Variable 2 : score de dépression -> quantitatif*
- *→ on compare des moyennes : test T de student pour séries appariées ou U de Mann et Whitney*
- *T_t = 2,26 (ddl = 10-1 = 9) et α = 5%)*
- *T_c = 2,91 > T_t*
- *Rejet de H0 au risque 5%*

Le placebo a un effet significatif

2. Le traitement est-il efficace ?

- *On compare les différences J28 – J0 de chaque patient, entre les 2 groupes*
- *H0 : il n'y a pas de différence entre le traitement et le placebo*
- *Variable 1 : traitement ou placebo -> qualitatif, Variable 2 : score de dépression -> quantitatif*
- *2 groupes indépendants de faibles effectifs → test T de student ou U de Mann et Whitney*
- *Dans la table, avec α = 5%, n_1 = 10 et n_2 = 10, u_t = 23*

Témoins d=J0-J28	3	2	-1	2	7	-1	2	3	3	1
Traités d=J0-J28	5	7	7	2	5	6	8	5	7	6

- *On classe ces différences par ordre croissant et on leurs associe un rang :*

-1	-1	1	2	2	2	2	3	3	3
1,5	1,5	3	5,5	5,5	5,5	5,5	9	9	9

5	5	5	6	6	7	7	7	7	8
12	12	12	14,5	14,5	17,5	17,5	17,5	17,5	20

Pour les valeurs en double, on calcule $\frac{\sum rangs}{nombre\ de\ valeurs}$. Par exemple pour -1 le rang est $\frac{1+2}{2}$ = 1,5. Pour 2 le rang est $\frac{4+5+6+7}{4}$ = 22,5.

- *On calcule u_1 : pour chaque témoin, on compte les traités classés avant :*
  u_1 = 0+0+0+0+0+0+1+1+1+6 = 9
- *On calcuel u_2 = 91 : soit on recalcule tout soit on sait que $u_1 + u_2 = n_1 * n_2$ donc 9 + u_2 = 10*10*
- *On prend u = min(u_1;u_2) = 9*
- *$U_c < U_t$: peu d'imbrication*
- *Rejet de H0 au seuil 5%*

Les différences sont significativement plus importantes avec le traitement qu'avec le placebo

***Conclusion** : le traitement est efficace contre la dépression*

Comment lire U_t dans la table ?

n_1

n_2-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
2	-	-	0	2	5	8	12	17	23	29
3	-	-	0	3	6	10	14	19	26	33
4	-	-	1	4	7	11	16	22	28	36
5	-	-	2	4	8	13	18	24	31	39
6	-	0	2	5	9	14	20	26	34	42
7	-	0	3	6	11	16	22	29	37	45
8	-	0	3	7	12	17	24	31	39	48
9	-	0	4	8	13	19	26	34	42	52
10	-	1	4	9	14	21	28	36	45	55
11	-	1	5	10	15	22	30	38	48	
12	-	1	5	11	17	24	32	41	50	
13	-	1	6	11	18	25	34	43		
14	-	1	6	12	19	27	36	45		
...										
18	-	2	8	16	24	33				
19	-	3	9	17	25					
20	-	3	9	17	27					

Ici c'est la table avec α = 5%

On regarde le plus petit des 2 effectifs sur les colonnes et la différence $n_2 - n_1$ sur les lignes

Ex : n_1 = 10 et n_2 = 10 : $n_2 - n_1$ = 0 donc U_t = 23

B. r' de Spearman

Ici le paramètre Z est r'

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

Si $r'_c > r'_t$ -> on accepte H0

Exemple : On prend la note de 6 étudiants en biostat et leur classement au concours PACES

X Biostat	*12,4*	*4,9*	*18,1*	*5,4*	*19,4*	*16*
Y Classement	*210*	*555*	*6*	*445*	*5*	*14*

H0 : il n'y a pas de lien entre ces 2 séries de valeurs numériques, il s'agit de 2 séries indépendantes

Variable 1 : note -> quantitative, Variable 2 : classement -> pseudo-quantitative

On associe à chaque X et à chaque Y un rang. On calcule d_i la différence entre le rang X et le rang Y et d_i^2

X Biostat	*12,4*	*4,9*	*18,1*	*5,4*	*19,4*	*16*
Rang X	*3*	*1*	*5*	*2*	*6*	*4*
Y Classement	*210*	*555*	*6*	*445*	*5*	*14*
Rang Y	*4*	*6*	*2*	*5*	*1*	*3*
d_i	*-1*	*-5*	*3*	*-3*	*5*	*1*
d_i^2	*1*	*25*	*9*	*9*	*25*	*1*

Dans la table, avec n = 6 et α = 5%, r'_t = 0,89. Avec α = 1%, r'_t = 1

r'_c = -1 < r'_t : on rejette H0

Il y a un lien significatif entre ces 2 séries. Plus la note de biostat est élevée, plus le classement est petit (d'où le signe – devant r'_c)

VI. Récap / Méthodologie d'utilisation des tests

Effectifs	Variables quantitatives	Variables qualitatives	Variables qualitative - quantitative
4<n<12 (non paramétrique)	**r' de Spearman**	Comparaison des pourcentages X^2	**U de Mann et Whitney**
12≤n<30	**Coefficient de corrélation** r' de Spearman	Comparaison des pourcentages X^2	**T de student** U de Mann et Whitney
30≤n	**Coefficient de corrélation** r' de Spearman	Comparaison des pourcentages X^2	**Comparaison des moyennes** T de student U de Mann et Whitney

On peut utiliser un test pour des effectifs supérieurs mais pas pour des effectifs inférieurs.

Remarque : le choix du test le plus approprié ne dépend pas que de l'effectif, il y a plein d'autres facteurs à prendre en compte (que l'on ne vous demande pas de connaître). Cela explique pourquoi le prof peut utiliser un test t de student avec un effectif de 10.

FIN