STATISTIQUES DÉDUCTIVES

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.
Exemple : pour comparer les notes à l'épreuve de biostatistiques entre deux années, on se pose la question : y a-t-il une différence entre ces deux groupes ?

Définition des hypothèses :

En statistiques descriptives on travaille à partir de 2 hypothèses :

Hypothèse H0 (ou hypothèse nulle)	Hypothèse H1 (ou hypothèse alternative)
► Il n'y a pas de différence entre les 2 groupes ► Les fluctuations observées sont dues au hasard	► Il existe une différence significative entre les deux groupes ► Les fluctuations observées ne sont pas dues au hasard

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

Étapes d'un test d'hypothèse :

1. Définir H0 et H1 *(à partir des données de l'énoncé)*
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 la paramètre Z
6. Utiliser la règle de rejet/acceptation de H0 : **comparer** le Z_c (paramètre calculé) au Z_t (paramètre théorique, dont on connaît la distribution)
7. Fixer le **risque d'erreur réel** (à posteriori)
8. Interpréter les résultats : interprétation statistique + médicale

Notion de risque :

Risque de première espèce / Risque α	Risque de seconde espèce / Risque β
► Probabilité de **rejeter** H0 si H0 est **vraie** ► Ce risque est **maîtrisé** ► Fixé à l'avance	► Probabilité **d'accepter** H0 si H0 est **fausse** ► Ce risque est **négligé** ► Fixé à posteriori ►Il peut être très élevé (en général ß = 20%)

Puisque le risque α est maîtrisé et le risque β est négligé,
il peut y avoir une **dissymétrie** dans le traitement des deux hypothèses.

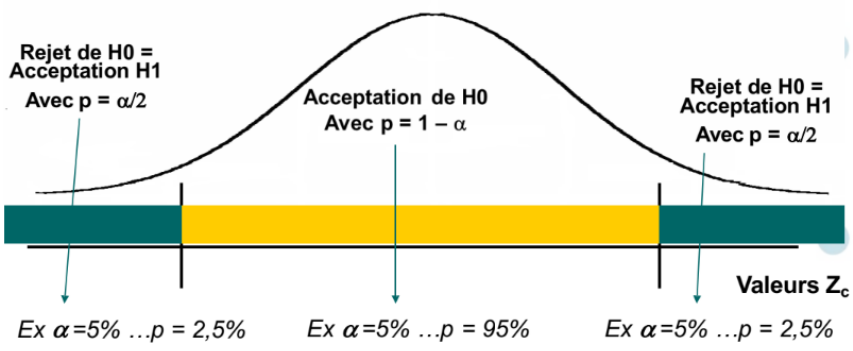
La puissance du test vaut 1 – ß
Elle correspond à la probabilité de rejeter H0 si H0 est fausse

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.**

		Décision du statisticiens	
		Rejet d'H0	Non rejet d'H0
Réalité	H0 vraie	Risque α	1 - α
	H0 fausse	1 – β (puissance du test)	Risque β

Interprétation graphique :

Le paramètre Z suit une distribution en forme courbe de Gauss (loi normale)



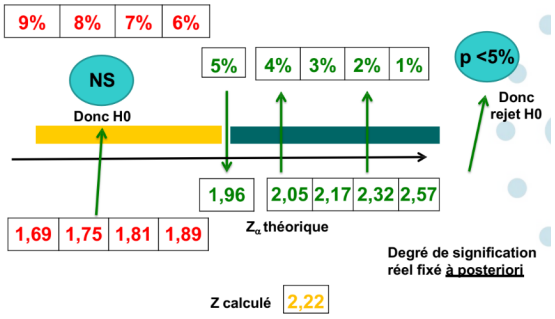
Pour arriver à une conclusion on doit :

1. Fixer le risque α à priori
2. Chercher Z_t dans la table
3. Calculer Z_c grâce aux formules
4. Comparer Z_c à Z_t ; on distingue deux situations :

$Z_c < Z_t$	$Z_c > Z_t$
Acceptation de H0 **p = 1 - α**	Rejet de H0 **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 supérieure à celle fixée au départ.



- *α = 5%*
- *Z_α = 1,96 (lu dans la table de l'écart réduit)*
- *Z_c = 2,22*
- *2,22 > 1,96 ($Z_c > Z_t$) donc on rejette H0*
- *Pour α = 1%, Z_α = 2,57, or 2,22<2,57*
($Z_c < Z_t$) donc on ne rejette pas H0 au risque 1%
- *On a donc p < 5%*

On pourrait dire qu'on rejette H0 à 3% (car 2,17<2,22<2,32, voir les chiffres en vert sur le schémas ci-dessus qui représentent Z_α), mais on ne le fait pas.

En pratique, on utilise seulement **1%** et **5%**. En effet, si on 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**	Situation **bilatérale**
Le rejet d'H0 permet seulement de dire qu'il y a une différence significative entre les 2 situations C'est la situation la plus **fréquente**	L'acceptation de H1 permet de déterminer laquelle des situations est la meilleure

Exemple : si on compare deux traitements A et B, en rejetant H0 :

- *en situation **unilatérale**, on pourra seulement 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 (ou inversement)*

Big data :

Et si les données étaient le pétrole du 21ème siècle ?
Nous générons et détenons quantités d'info personnelles : alimentation, achats, contributions aux 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 deux 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 ?**).
Ils 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**.

LIEN ENTRE DEUX VARIABLES **QUALITATIVES**

On se demande si le pourcentage d'individus 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, …)

Test de comparaison des pourcentages (tout effectif) :

Le paramètre Z est l'écart réduit ε

- ► ε_t vient de la table de l'écart réduit
- ► $$\varepsilon_c = \frac{p_A - p_B}{\sqrt{\frac{p_A q_A}{n_A} + \frac{p_B q_B}{n_B}}}$$

  $q_A = 1 - p_A$
  p = probabilité d'être malade
  n : taille de l'échantillon
- ► **Si $\varepsilon_c > \varepsilon_t$ → rejet de H0**

Méthodologie pour chercher ε_t dans la 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

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	3,89059	4,41717	4,89164	5,32672	5,73073	6,10941

On cherche ε_t en fonction d'α
*On regarde le **dixième d'α sur les lignes** et le **centième sur les colonnes**. ε_t sera à l'intersection.*
E.g : Pour α = 5% = 0,05 : on regarde 0,00 dans les lignes et 0,05 dans les colonnes : $\varepsilon_t = 1,96$
Pour α = 0,1% = 0,001 : on regarde la table des petites valeurs $\varepsilon_t = 3,29$

***Exemple** : Soient 2 groupes de 200 enfants : Crèche : 200 enfants, 130 cas de rhinopharyngite*
Maison : 200 enfants, 96 cas de rhinopharyngite

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

1. *H0 : pas de différence entre les 2 modes de garde vis-à-vis du développement de rhino*
   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 de pourcentages
3. *α = 5%, défini à priori*
4. $p_A = 130/200$ $p_B = 96/200$
   $p_A = 65\%$ $p_B = 48\%$
   $\varepsilon_c = \dfrac{0{,}65 - 0{,}48}{\sqrt{\dfrac{0{,}65 \times 0{,}35}{200} - \dfrac{0{,}48 \times 0{,}52}{200}}}$ *(vous n'aurez pas à reproduire ce calcul, le paramètre calculé vous sera donné)*

5. $\varepsilon_c = 3{,}4$
   ***3,4 > 3,3 donc on rejette H0 au seuil 0,001** (3,3 vient de la table de l'écart-réduit pour les petites valeurs)*
   On a donc $p \leq 0{,}001$

Sur cet échantillon, le risque de rhinopharyngite 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, …), d'où l'importance de distinguer l'interprétation statistique et médicale …

Test du X^2 (Tout effectif) :

On utilise **de préférence** ce test si notre tableau de données a plus de 2 lignes (ou 2 colonnes)
Le paramètre Z est $\mathbf{X^2}$

- X^2_t vient de la table du X^2
- $X^2_c = \Sigma \dfrac{(o_i - c_i)^2}{c_i}$ 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) x (nombre de colonnes – 1)**

Exemple : exposition au benzène et 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%*
4. *Valeurs observées : 15, 20, 485 et 980*
   Valeurs calculées (obtenues par un modèle théorique) :
   ● il y a 35 malades pour 1500 personnes au total soit 2,33% de malade.
   On applique 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 exposés et aux non- exposés :
   - 97,67% de 500 = 488,3
   - 97,67% de 1500 = 976,7

$$X_c^2=\frac{(15-11{,}65)^2}{11{,}65}+\frac{(20-23{,}35)^2}{23{,}35}+\frac{(485-488{,}3)^2}{488{,}3}+\frac{(980-976{,}7)^2}{976{,}7}=1{,}42$$

*ddl = (2-1) * (2-1) = 1 donc X_t^2 = 3,84*

5. *$X_c^2 < X_t^2$ donc on accepte H0 au seuil 0,05*
   Il n'existe pas de relation entre l'exposition au benzène et les leucémies

Table du X^2

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 valeurs nécessaires dans une série pour pouvoir calculer toutes les autres.
On cherche le ddl sur les lignes et α sur les colonnes

LIEN ENTRE VARIABLES **QUALITATIVES** ET **QUANTITATIVES**

En moyenne, la taille des individus d'une population A coïncide-t-elle avec la taille des individus d'une population B ?

Test de comparaison de moyennes (**n_1 et n_2 > 30 : grands échantillons**) :

Le paramètre Z est l'écart-réduit ε

- ► ε_t vient de la table de l'écart-réduit
- ► $\varepsilon_c = \dfrac{m_1 - m_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$
- ► **Si $\varepsilon_c > \varepsilon_t$ → rejet de H0**

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

Test T de student (**n_1 ou n_2 < 30 : petits échantillons**) :

Le paramètre Z est t
t_t : lu dans la table du t de student

- ► $s = \sqrt{\dfrac{\Sigma(x_i - m_1)^2 + \Sigma(x_j - m_2)^2}{(n_1 - 1) + (n_2 - 1)}}$ (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)**

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.

Précision sur le ddl *: nombre minimal de valeurs d'une série, nécessaire afin de pouvoir calculer les manquants si l'on dispose du total ou des totaux des valeurs de cette série.*
(n est le nombre de valeurs par ligne, comme le nombre de notes dans un semestre par exemple)

2	*3*	*5*	*12*	*10*	*x*	*7*	*8*	*Tot : 51*
2	*3*	*5*	*12*	*10*	*y*	*z*	*8*	*Tot : 51*

Avec n-1 valeur et le total, on peut trouver que x=51-2-3-5-12-10-7-8
Avec n-2 valeurs, on ne peut pas trouvez les deux valeurs manquantes
Le degré de liberté est donc de n-1 ici.

Exemple *: Soient 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 ?*
n1 = 15 ; m 1 = 6,3 ; s1 = 1,8
n2 = 12 ; m2 = 4,5 ; s2 = 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 *et* $n_2 < 30$ *: test T de student*
3. *a = 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*

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 autre à ce même groupe 6 mois plus tard. Il y a un lien entre les premiers et les derniers résultats car l'analyse sanguine est propre à chacun.*

Si **n > 30** on utilise le test de comparaison des moyennes : $\varepsilon = \frac{m_d}{\sqrt{\frac{s^2}{n}}}$

Si **n < 30** on utilise le test t de student : $t = \frac{m_d}{\sqrt{\frac{s^2}{n}}}$

Le reste de la méthodologie est identique.

d : différence de résultat pour un même sujet
m_d *: moyenne des d*
s : variance des d
n : nombre de couples

***Exemple** : On souhaite évaluer l'effet d'une substance S capable de désintoxiquer les fumeurs. On confectionne 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.*

1. Quelle est la première précaution à prendre ?
Les 2 groupes doivent être comparables pour les paramètres qui peuvent influencer le traitement : âge, sexe, CSP (catégorie socio-professionnelle), consommation par jour.

On compare donc les consommations moyennes avant traitement 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 à test de comparaison des moyennes
- $\varepsilon_c = 2 > 1{,}96$ *: Rejette de H0 avec un risque α = 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 traitement ?
- *Variable 1 : avant après ttt = qualitatif*
  Variable 2 : C = quantitative.
  Échantillon non indépendants à méthode des couples ; n > 30 à 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 traitement ?
- *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 traitement ?
Il faut comparer les variations dans les 2 groupes pour prouver l'efficacité de la substance S
1. *H0 : il n'existe pas de différence entre les variations de consommation dans les 2 groupes*
2. *Variable 1 : S ou P = qualitative*
   Variable 2 : C = quantitative
   n > 30 à test de comparaison des moyennes
3. $\varepsilon_c = 2{,}09 > 1{,}96$
4. *Rejet de H0 au risque 5%*
5. *Il existe une différence significative entre les variations de consommation dans les 2 groupes (p<5%)*
6. *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 dire qu'il y a une efficacité du traitement S pour désintoxiquer les fumeurs.*

LIEN ENTRE DEUX VARIABLES **QUANTITATIVES**

Corrélation et régression :

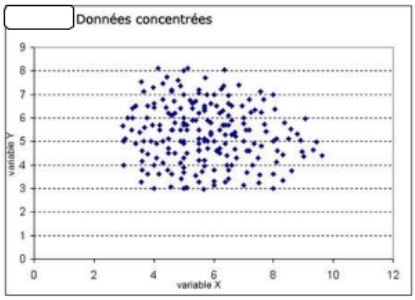
Corrélation : évaluation de la liaison entre 2 variables quantitatives
Régression : méthode mathématique permettant d'expliquer les relations entre les variables observées

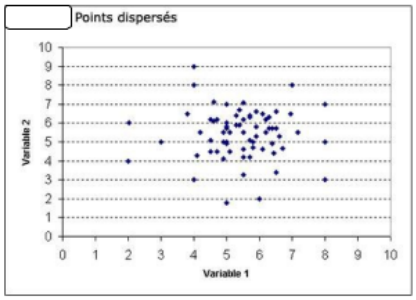
Représentation des données :

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

Nuages de points :

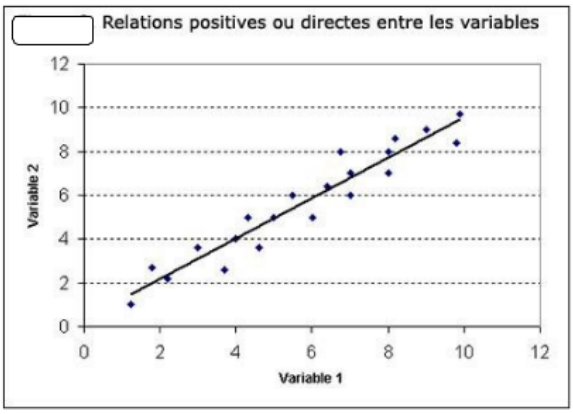
Il n'y a pas de relation entre x et y :





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 (en forme de cloche)

► Si x et y sont liés alors y = f(x), et on obtient une **droite de régression de y en x**



Étude de la liaison entre caractères quantitatifs :

Exemple : 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** certaines valeurs 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 recueille leur âge et leur concentration de cholestérol :*

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

Le taux de cholestérol est-il lié à l'âge ?

- *H0 : le taux de cholestérol n'est pas lié à l'âge*
  Variable 1 : Age = quantitatif
  Variable 2 : taux de cholestérol = quantitatif
- *→ Test du coefficient de corrélation*
- *α = 1%,*
- *DDL = 10-2 = 8 donc r_t = 0,76 (lu dans la table du coefficient de corrélation, à l'intersection entre ddl et α)*
- *r'_c = 0,955 > r'_t : Rejet de 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*
On peut des faire de corrélations de tout et n'importe quoi, on peut tracer des courbes qui montrent une relation de proportionnalité sans pour autant qu'x influe y. C'est le rôle des statistiques et des essais cliniques de déterminer si ce lien de corrélation est un lien de causalité ou non.

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.

U de Mann et Whitney :

Le test 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 teste donc la liaison entre une variable qualitative et quantitative).

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 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 : $u = \min\{u1 ; u2\}$
7. On compare u_C 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_α tel que $P(U \leq m_\alpha) = \alpha$

► On rejette H0 au risque α si $\mathbf{u \leq m_\alpha}$, 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

Y a-t-il un effet placebo ?

1. *H0 : le placebo n'a aucun effet, les scores J0 ne diffèrent pas des scores J28*
2. *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
3. *T_t = 2,26 (ddl = 10-1 = 9) et a = 5%)*
4. *T_c = 2,91 > Tt*
5. *Rejet de H0 au risque 5%. Le placebo a un effet significatif.*

Le traitement est-il efficace ?

On compare les différences J28 – J0 de chaque patient, entre les 2 groupes :

1. *H0 : il n'y a pas de différence entre le traitement et le placebo*
2. *Variable 1 : traitement ou placebo -> qualitatif*
   Variable 2 : score de dépression -> quantitatif
3. *2 groupes indépendants de faibles effectifs à test T de student ou U de Mann et Whitney*
4. *Dans la table, avec a = 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 :

Valeurs obtenues	-1	-1	1	2	2	2	2	3	3	3	5	5	5	6	6	7	7	7	7	8
Rang associé	1,5	1,5	3	5,5	5,5	5,5	5,5	9	9	9	12	12	12	14,5	14,5	17,5	17,5	17,5	17,5	20

- *Pour les valeurs en double, on calcule* $\frac{\Sigma_{rang}}{nombre\ de\ valeurs}$
  Par exemple pour -1 le rang est (1+2)/2 = 1,5.
  Pour 2 le rang est (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 calcule u_2 = 91 : soit on recalcule tout, soit on sait que $u_1 + u_2 =$ 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, le traitement est efficace contre la dépression.*

*

Table U de Mann et Whitney (α = 5%)

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, n1 = 10
n2 – n1 =
10 – 10 = 0
D'où U_t = 23

r' de Spearman :

On étudie la liaison entre deux caractère quantitatifs pour de **faibles échantillons**.
Le paramètre Z est r'.

- $r' = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)}$
- **Si $r'_c > r'_t$ -> on ACCEPTE H0**

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

X : note	*12,4*	*4,9*	*18,1*	*5,4*	*19,4*	*16*
Y : classement	*210*	*555*	*6*	*445*	*5*	*14*

1. *H0 : il n'y a pas de lien entre ces 2 séries de valeurs numériques, il s'agit de 2 séries indépendantes*
2. *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, puis on calcule* 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 = 3 - 4*	*-5 = 1 - 6*	*3*	*-3*	*5*	*1*
d_i^2	*1 = (-1)²*	*25 = (-5)²*	*9*	*9*	*25*	*1*

Dans la table, avec n = 6 et a = 5%, $r'_t = 0{,}89$ *(lu dans la table, à l'intersection de alpha et n)*
pour 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 diminue (d'où le signe – devant r'_c *)*

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

En gras : les test les plus appropriés

Remarque : 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 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.
Remarque (explicitée par le professeur à la séance de réponse aux question il y a 3 ans) : les formules pour calculer les paramètres de chaque test ne sont pas à savoir refaire, sauf celle du X2. Les résultats seront donc donnés dans l'énoncé s'ils sont nécessaires pour répondre aux questions.

***Remarque personnelle** : Le raisonnement du test de U de Mann et Whitney ou du X^2 ne sera pas à refaire entièrement, du moins pas tant qu'on aura 1 minute par qru.*
***Remarque personnelle** : pour déterminer quel test est le plus approprié pour un test, ne prenez pas en compte la remarque selon laquelle on peut utiliser un test pour des effectifs supérieurs. Attention, si le test le plus approprié n'est pas demandé, on peut citer les autres « moins appropriés ».*
De manière générale, le test le plus adéquat était directement précisé dans l'énoncé des annales, pour éviter toute confusion. Le professeur est conscient que ces notions sont délicates à manipuler, et ne vous piégera pas du des notions non abordées dans ce cours.

Dédicaaaaaaaaces !!!!!!
Dédi aux gens qui lisent les dédi avant de lire cours (comme moi)
Dédi au tutorat niçois, qui me donne espoir en l'humanité
Dédi à mon chat, qui a failli me faire rater l'année à force de me déconcentrer
Dédi aux gens qui disent bonjour dans la rue
Dédi aux gens qui disent bonjour tout cours
Dédi à Baqué qui nous a tous trollé à l'examen
Dédi à mon addiction aux cacahuètes
Dédi à tous les addicts de la buv
*Et enfin dédi à moi, pour avoir codé manuellement sur LibreOffice toutes les ********** de formules de ******** car je pouvais pas copier-coller depuis word -_-*