Tut-rentrée 2015-2016

Cours 3-UE4-Biostatistiques

Statistique Descriptive

- Description de populations
- Calculs d'estimateurs
- Notion de sondages

Les paramètres

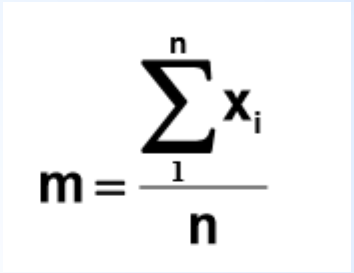
Les paramètres permettent de « résumer » les caractéristiques de la série statistique

- Seulement pour des variables **quantitatives**

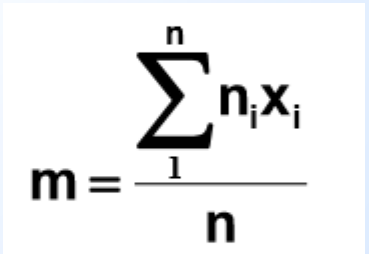
Indicateurs de **position**	Indicateurs de **dispersion**
Moyenne	Variance
Médiane	Ecart-type
Quartiles	*Poutoux* ♡

La Moyenne

- Variable quantitative discrète :



- Variable quantitative continue :



La médiane

- Ordonner par ordre croissant
- Séparer en 2 groupes de même effectif (50% au dessus 50% au dessous)

- Effectif n pair :

$$m = \frac{x_{n/2} + x_{(n/2)+1}}{2}$$

- Effectif n impair

$$m = \frac{x_{n+1}}{2}$$

Exemple : 3,4,**6**,8,10

N=5

Moyenne VS Médiane

- **La Moyenne** :
- Facile à calculer
- Très utilisée dans les tests statistiques
- Significative si répartition des données symétrique et dispersion faible
- **Sensible aux valeurs anormales**

- **La Médiane** :
- Calcul facile, peu sensible aux valeurs anormales
- Valeurs ordinales
- **Peu adaptée aux séries statistiques**

Les Quartiles

Les quartiles partagent la série ordonnée en 4 groupes de même effectif

- **Q1** (premier quartile) sépare les premiers 25% de la série
- **Q2** (deuxième quartile = **Médiane**) sépare les premiers 50% de la série
- **Q3** (troisième quartile) sépare les premiers 75% de la série

!Point méthode!

Calcul des Quartiles

- On range les données par ordre croissant
- Si N (effectif total) est un **multiple de 4** :

☆ Q1 = N/4
☆ Q2 = Médiane
☆ Q3 = 3xN/4

- Si N n'est **pas** un multiple de 4 :

☆ Q1 = Ni+Nj/2 *avec i et j : i<N/4<j*
☆ Q3 = Nk+Nl/2 avec k<3N/4<l

La Variance

Paramètre indiquant la dispersion des données autour de la moyenne

- Variance = (Ecart-type)²

L'Ecart-type

« Moyenne de l'écart à la moyenne »

QCMs

- Qcm 1 :

A) La moyenne est un indicateur de dispersion

B) La variance est un indicateur de position

C) Le Quartile est un indicateur de position

D) L'Ecart-type est un indicateur de dispersion

E) Toutes les réponses sont fausses

QCMs

- Qcm 1 :

A) La moyenne est un indicateur de dispersion

B) La variance est un indicateur de position

C) Le Quartile est un indicateur de position

D) L'Ecart-type est un indicateur de dispersion

E) Toutes les réponses sont fausses

Numéro du Pokemon	1	2	3	4	5	6	7
Taille	125	118	121	122	121	121	124

- Moyenne ?
- Médiane ?
- Q1/Q2/Q3 ?

- Moyenne → 121,7 cm
- Médiane → 118, 121, 121, **121**, 122, 124, 125
- Q1 → 7/4 = 1,75 : on fait la moyenne de la 1ère et 2ème valeur : Q1= 119,5
- Q2 = MEDIANE = 121
- Q3 → 3x7/4 = 5,25 : on fait la moyenne entre la 5ème et 6ème valeur : Q3 = 123

L'estimation statistique

- Estimation **PONCTUELLE** : Valeur, jugée la meilleure à un instant t (peu fiable)
- Estimation par **INTERVALLE** : Intervalle de valeurs contenant la valeur recherchée = intervalle de confiance ♡♡♡

Méthodologie :

- Détermination précise de la population étudiée = population cible
- Tirage au sort n sujets → Echantillon représentatif
- Calcul de l'intervalle de confiance

- Deux **estimations ponctuelles** d'une même variable réalisées sur les échantillons A et B donneront des **valeurs ponctuelles voisines**, mais pas nécessairement les mêmes valeurs.

- Deux **estimations par intervalles** d'une même variable réalisées sur les échantillons A et B donneront des **Intervalles de confiance (IC) qui se recouvrent**, mais pas nécessairement les mêmes.

L'Intervalle de confiance

$$\mu \in [\, m \pm \frac{\varepsilon s}{\sqrt{n}} \,] \quad \Rightarrow \quad \text{Intervalle au risque } \alpha$$

A connaître :

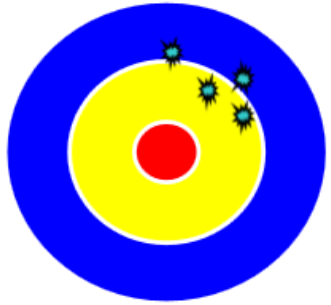
$\alpha = 5\% \rightarrow \varepsilon = 1.96$

$\alpha = 1\% \rightarrow \varepsilon = 2.6$

Si α diminue ε augmente

Précision de l'estimation

Large = plus de chances de l'atteindre, mauvaise précision de l'estimation



Resserré = risque de rater, meilleure précision de l'estimation

$$i = \varepsilon \frac{s}{\sqrt{n}}$$

$$n = \varepsilon^2 \frac{s^2}{i^2}$$

ATTENTION : Plus l'indice est petit meilleure est la précision

$$\mu \in [\, m \pm \frac{\varepsilon s}{\sqrt{n}} \,] \Rightarrow \text{Intervalle au risque } \alpha$$

- Plus α est petit plus l'intervalle est grand
- Si la taille de l'échantillon augmente, la précision augmente
- Plus l'IC est large, moins il est précis

QCM Time

- QCM 1 : On tire au sort un échantillon de 100 étudiants sur lequel on cherche à étudier le niveau de concentration lors d'un cours de biostat' sur une échelle de 0 à 10.

On donne : Moyenne m= 7 ; Ecart-type = 1. On travaille au risque α = 5%.

A) L'intervalle de confiance est de $[7 \mp 1{,}96 \times 1/10]$

B) L'intervalle de confiance est de $[7 \mp 2{,}6 \times 1/10]$

On décide d'augmenter l'échantillon à 10 000 étudiants :

C) La taille de l'intervalle de confiance diminue d'un facteur 100

D)L'indice de précision diminue donc la précision diminue

E) Le niveau de concentration est toujours de 10

QCM Time

- QCM 1 : Soit un échantillon de 100 étudiants sur lequel on cherche à étudier le niveau de concentration lors d'un cours de biostat' sur une échelle de 0 à 10.

On donne : Moyenne m= 7 ; Ecart-type = 1. On travaille au risque α = 5%.

A) L'intervalle de confiance est de $[7 \mp 1{,}96 \times 1/10\ ]$

B) L'intervalle de confiance est de $[7 \mp 2{,}6 \times 1/10\ ]$

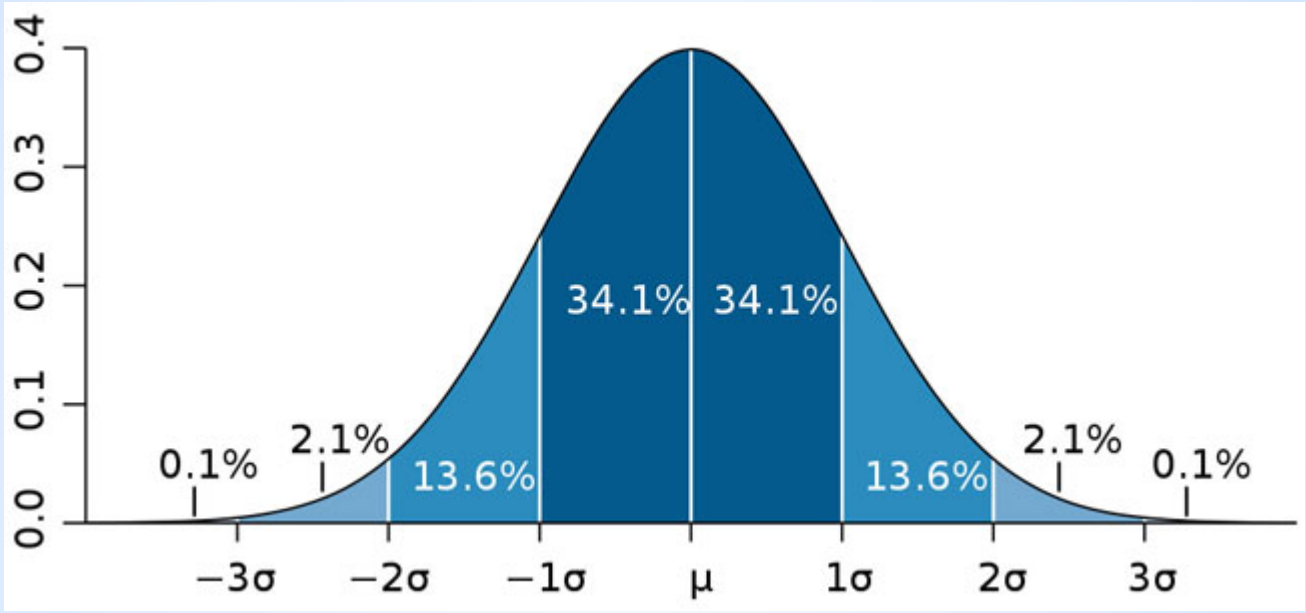
On décide d'augmenter l'échantillon à 10 000 étudiants :

C) La taille de l'intervalle de confiance diminue d'un facteur 100

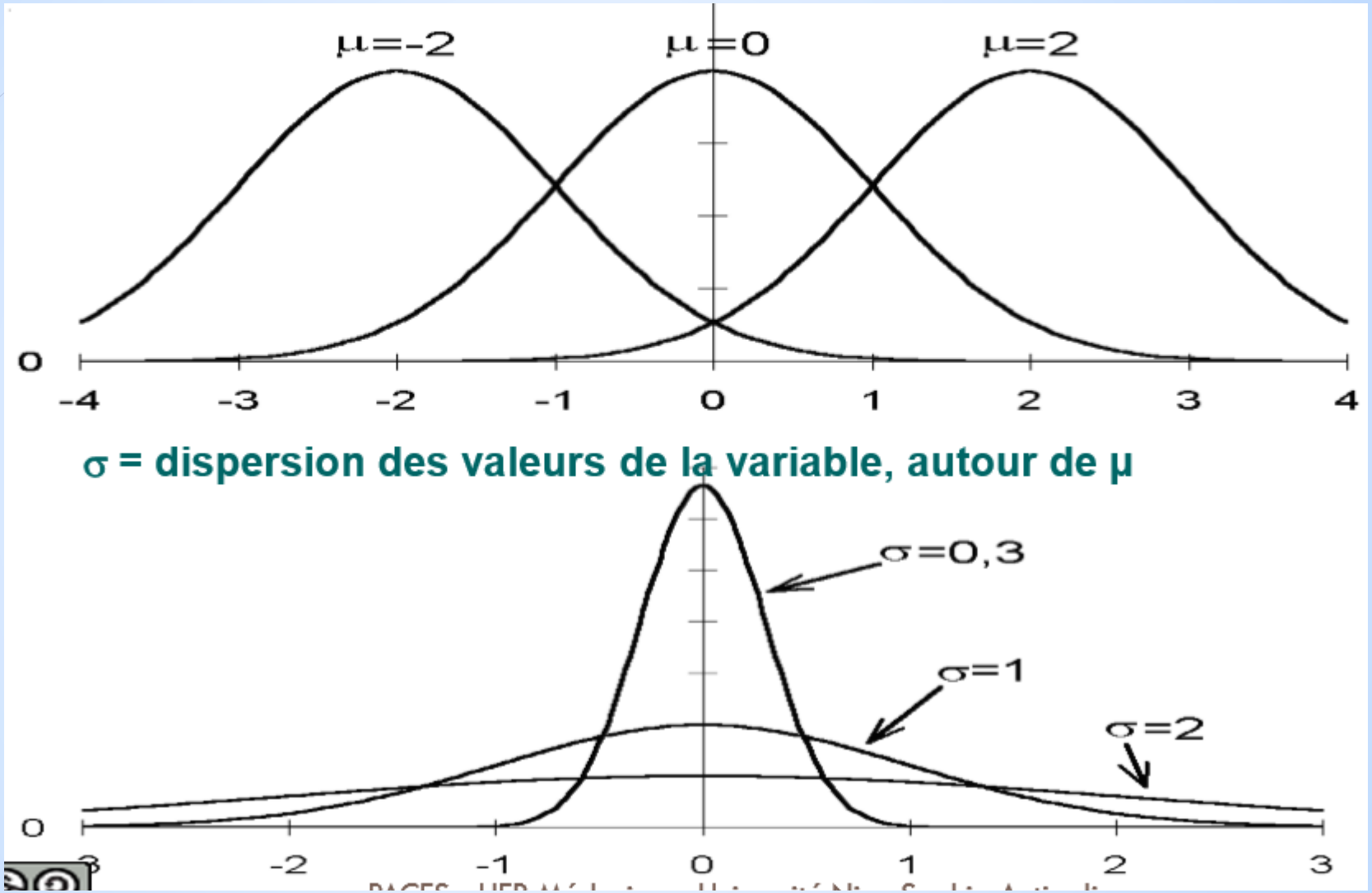
D) L'indice de précision diminue donc la précision diminue

E) Le niveau de concentration est toujours de 10 ☹

La lois Normal ou Courbe de Gauss



- Entre (m-s) et (m+s) on a 68.2% de la population
- Entre (m-1.96s) et (m+1.96s) on a 95.4% de la population
- Entre (m-2.6s) et (m+2.6s) on a 99.6% de la population



Estimation de données qualitatives

ECHANTILLON	POPULATION TOTALE
n = effectif	N = effectif
Po = pourcentage observé	p = pourcentage réel
s = ecart type	σ = ecart type

Estimation de l'écart-type inconnu :

$$s = \sqrt{\frac{p_0 q_0}{n}}$$

avec $q_0 = 1 - p_0$

Intervalle de confiance

$$p \in [\, p_{obs} - \varepsilon s \; ; \; p_{obs} + \varepsilon s \,]$$

Indice de précision des données qualitatives : $i = \varepsilon.s = \varepsilon. \sqrt{po.qo/n}$

$$i = \varepsilon \sqrt{\frac{p(1-p)}{n}}$$

$$n = \varepsilon^2 \frac{p(1-p)}{i^2}$$

Statistiques Déductives

*Tirer des **conclusions** à partir d'**observations***

2 **Hypothèses** :

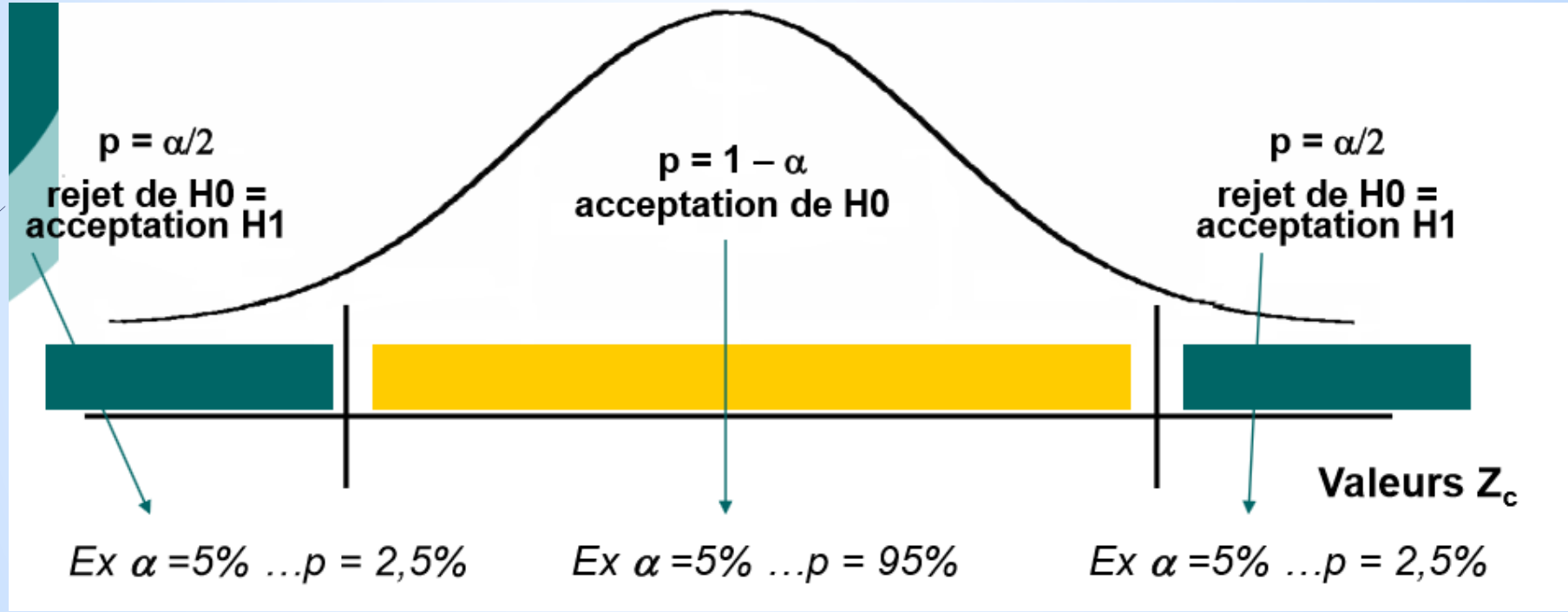
- H0 = Hypothèse nulle. Pas de différence observée
- H1 = Hypothèse alternative. Différence significative entre les 2 groupes

Les **Tests** : Techniques permettant de décider si on garde ou repousse H0, en ayant fixé le risque d'erreur accompagnant cette décision

On choisit toujours pour **H0** l'hypothese qu'il serait le plus grave de rejeter a tort

Les étapes de mise en œuvre des tests d'hypothèse

- Etape 1 : Avant recueil des données définir H0 et H1. Les 2 hypothèses jouent des rôles symétriques
- Etape 2 : Avant recueil des données définir le test en fonction du type des données (qualitatives, quantitatives). Soit Z le paramètre qui sera calculé
- Etape 3 : Avant recueil des données on choisi le risque a (dans la pratique souvent 5%)
- Etape 4 : Recueil des données + Calcul de Z
  Règle de décision : examiner la position de cette valeur Z, par rapport à un modèle théorique dont on connait la distribution.
- Etape 5 : Interprétation des résultats.



Notion de risques

- Risque de première espèce : α Probabilité de rejeter H0 si H0 vraie.
  compromis universel : $\alpha = 5\%$
- Risque de seconde espèce : β Probabilité d'accepter H0, si H0 fausse
- Puissance d'un test : $1 - \beta$ Probabilité de rejeter H0 si H0 fausse

Réalité	Décision du statisticien : Rejet H0	Décision du statisticien : Non rejet H0
H0 Vraie	Erreur 1ère espèce α	$1 - \alpha$
H1 Vraie	Puissance $1 - \beta$	Erreur 2ème espèce β

QCMS

- QCM 1 :

A) Le risque α revient à condamner un innocent

B) β représente la puissance du test

C) Le risque de 2nde espèce revient à rejeter une différence qui existe

D) H0 est l'hypothèse nulle

E) Toutes les réponses sont fausses

QCMS

- QCM 1 :

A) Le risque α revient à condamner un innocent

B) β représente la puissance du test : c'est 1- β

C) Le risque de 2nde espèce revient à rejeter une différence qui existe

D) H0 est l'hypothèse nulle

E) Toutes les réponses sont fausses

LES TESTS ♡

Etude de la liaison entre 2 caractères qualitatifs

1) Test de comparaison des pourcentages

$$\varepsilon_{\text{(calculé)}} = \frac{p_A - p_B}{\sqrt{\frac{p_A q_A}{n_A} + \frac{p_B q_B}{n_B}}}$$

N'est pas à apprendre

- Si ε calculé > ε théorique : on accepte H1 et on rejette H0 : il y a une différence
- Si ε calculé < ε théorique : on accepte H0 : il n'y a pas de différence

Nombre de degrés de libertés

Le nombre de degré de liberté (ou ddl) se traduit par le nombre minimal de données qu'il est nécessaire de connaître afin de pouvoir déduire toutes les données manquantes.

- leur somme =0
- il suffit d'en connaitre (n-1) pour les connaitre tous : n-1 degrés de liberté

- Exemple :

X+Y = 18
Si Y = 12 alors X vaut obligatoirement 6
→ 2 variables pour 1 degrés de libertés

Etude de la liaison entre 2 caractères qualitatifs

2) Test du c^2

C^2 calculé = $\sum \frac{(Oi - Ci)^2}{Ci}$ Pas à apprendre

Le c^2 théorique est donné par la table du c^2
Pour le test du c^2 , le nombre de ddl vaut : (nblignes – 1) x (nbcolonnes – 1)

❖Si c^2 calculé > c^2 théorique : on rejette H0 : on accepte H1 au risque α (on conclu qu'il y a une différence)

❖Si c^2 calculé < c^2 théorique : on accepte H0 au risque α (il n'y a pas de différence)

Table du c^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
...									

Liaison entre caractères qualitatifs et quantitatifs

1) Comparaison des Moyennes

- n1 et n2 > 30 : grands échantillons
- Table de l'écart réduit

2) Test t de student

- n1 ou n2 < 30 : petits échantillons
- Table du t de student
- Nb ddl : (n1 – 1) + (n2 - 1)

Table du t de student

α / ddl	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	**63,656**	636,578
2	0,142	0,816	1,386	1,886	2,920	**4,303**	6,965	**9,925**	31,600
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	1,383	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
12	0,128	0,695	1,083	1,356	1,782	**2,179**	2,681	**3,055**	4,318
13	0,128	0,694	1,079	1,350	1,771	**2,160**	2,650	**3,012**	4,221
14	0,128	0,692	1,076	1,345	1,761	**2,145**	2,624	**2,977**	4,140
15	0,128	0,691	1,074	1,341	1,753	**2,131**	2,602	**2,947**	4,073
16	0,128	0,690	1,071	1,337	1,746	**2,120**	2,583	**2,921**	4,015
17	0,128	0,689	1,069	1,333	1,740	**2,110**	2,567	**2,898**	3,965
18	0,127	0,688	1,067	1,330	1,734	**2,101**	2,552	**2,878**	3,922
19	0,127	0,688	1,066	1,328	1,729	**2,093**	2,539	**2,861**	3,883
20	0,127	0,687	1,064	1,325	1,725	**2,086**	2,528	**2,845**	3,850
21	0,127	0,686	1,063	1,323	1,721	**2,080**	2,518	**2,831**	3,819
22	0,127	0,686	1,061	1,321	1,717	**2,074**	2,508	**2,819**	3,792
23	0,127	0,685	1,060	1,319	1,714	**2,069**	2,500	**2,807**	3,768
24	0,127	0,685	1,059	1,318	1,711	**2,064**	2,492	**2,797**	3,745
25	0,127	0,684	1,058	1,316	1,708	**2,060**	2,485	**2,787**	3,725
26	0,127	0,684	1,058	1,315	1,706	**2,056**	2,479	**2,779**	3,707
27	0,127	0,684	1,057	1,314	1,703	**2,052**	2,473	**2,771**	3,689
28	0,127	0,683	1,056	1,313	1,701	**2,048**	2,467	**2,763**	3,674
29	0,127	0,683	1,055	1,311	1,699	**2,045**	2,462	**2,756**	3,660
30	0,127	0,683	1,055	1,310	1,697	**2,042**	2,457	**2,750**	3,646
40	0,126	0,681	1,050	1,303	1,684	**2,021**	2,423	**2,704**	3,551
80	0,126	0,678	1,043	1,292	1,664	**1,990**	2,374	**2,639**	3,416
120	0,126	0,677	1,041	1,289	1,658	**1,980**	2,358	**2,617**	3,373
∞	0,126	0,675	1,037	1,282	1,645	**1,960**	2,327	**2,577**	3,293

Liaison entre caractères quantitatifs

- **<u>Le coefficient de corrélation</u>** :

- r est compris entre [-1;1] avec n – 2 ddl
- Si r > 0 liaison positive : x et y varient dans le même sens
- Si r < 0 liaison négative : x et y varient en sens inverse

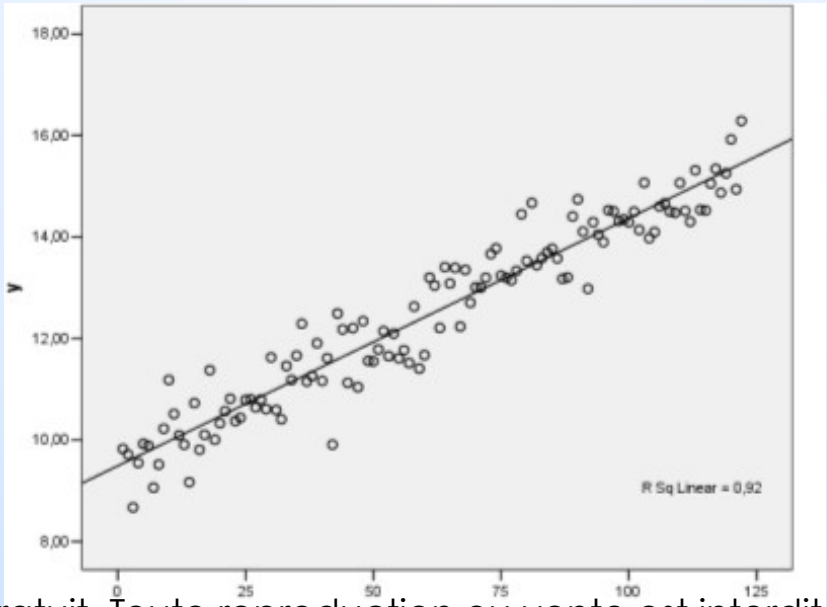


Table du coefficient de corrélation

ddl	0,1	0,05	0,02	0,01
1	0,9877	0,9969	0,9995	0,9999
2	0,9	0,95	0,98	0,99
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	0,576	0,6581	0,7079
11	0,4762	0,5529	0,6339	0,6835
12	0,4575	0,5324	0,612	0,6614
13	0,4409	0,5139	0,5923	0,6411
14	0,4259	0,4973	0,5742	0,6226
15	0,4124	0,4821	0,5577	0,6055
16	0,4	0,4683	0,5425	0,5897
17	0,3887	0,4555	0,5285	0,5751
18	0,3783	0,4438	0,5155	0,5614
....				
100	0,1638	0,1946	0,2301	0,254

QCMS time

- QCM 1 : On cherche a savoir si la place dans l'amphi influence l'apparition d'escarres fessiers : On réalise une étude statistique :

εcalculé= 2,1

A) On peut utiliser le test du c^2

B) On peut utiliser le test de comparaison des %

C) Au risque $\alpha = 5\%$ on accepte H0

D) Au risque $\alpha = 1\%$ on accepte H1

QCMS time

- QCM 1 : On cherche a savoir si la place dans l'amphi influence l'apparition d'escarres fessiers : On réalise une étude statistique :

εcalculé= 2,1

A) On peut utiliser le test du c^2 : comparaison de 2 variables qualitatives

B) On peut utiliser le test de comparaison des pourcentages

C) Au risque $\alpha = 5\%$ on accepte H0 εthéorique pour $\alpha = 5\%$ est de 1,96 donc εthéorique< εcalculé on accepte H1 au risque 5%

D) Au risque $\alpha = 1\%$ on accepte H1 εthéorique (=2,6) >εcalculé on accepte H0

FIN

Un P1 après 2h de biostats