UE 4:

BIOSTATISTIQUES

FIN STATISTIQUE DESCRIPTIVE + INTRO STAT DÉDUCTIVE.

PR. BENOLIEL/SKIINI

Rappel:

α (risque de 1ere espece)= 5 % => ε (ecart réduit) =1,96

Pour α = 1 % => ε=2,6

:

Pour le nombre de sujet necessaire n'apprenez pas par cœur la formule.. !!

Vous avez juste besoin de l'indice de précision:

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

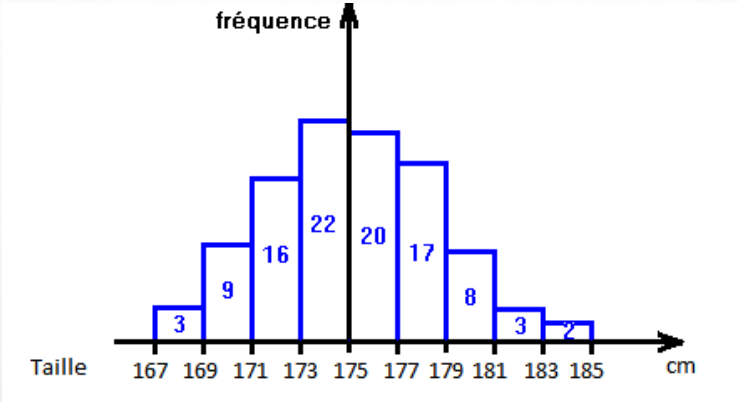
$$i^2 = \varepsilon^2 \frac{s^2}{n}$$

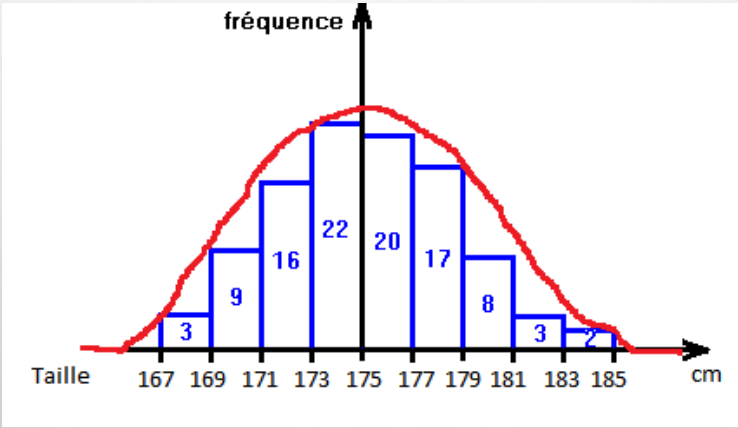
Donc $n = \varepsilon^2 \frac{s^2}{i^2}$

La loi normale ou de Gauss (en cloche) :

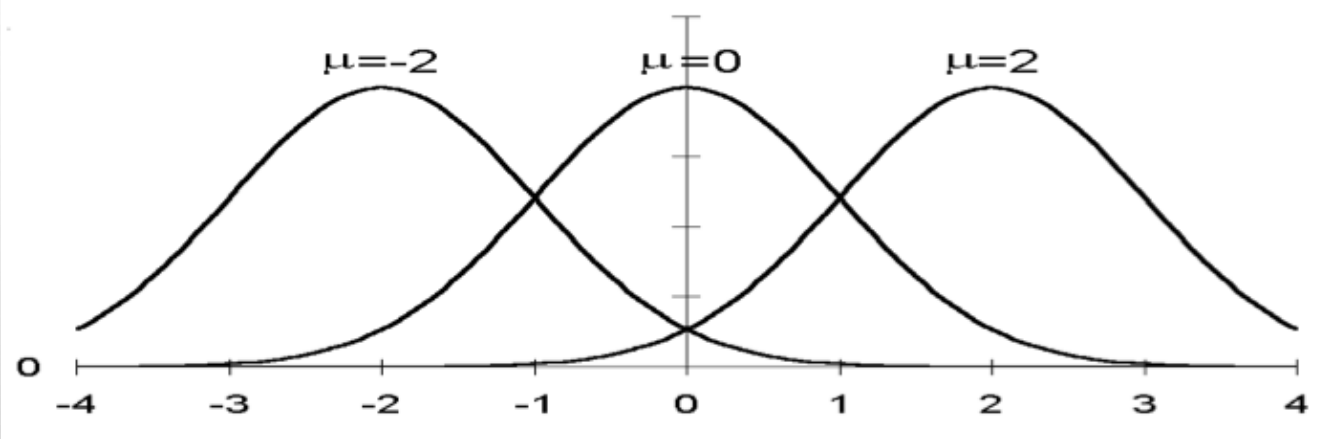
Je mesure la taille d'un certain nombre d'homme. Si on fait **tendre** le nombre de mesures vers l'**infini**, cette courbe **tend** vers une courbe **limit**e appelée « **courbe de Gauss** » :

Applicable que pour des populations de **plus de 30** personnes !!!!!!



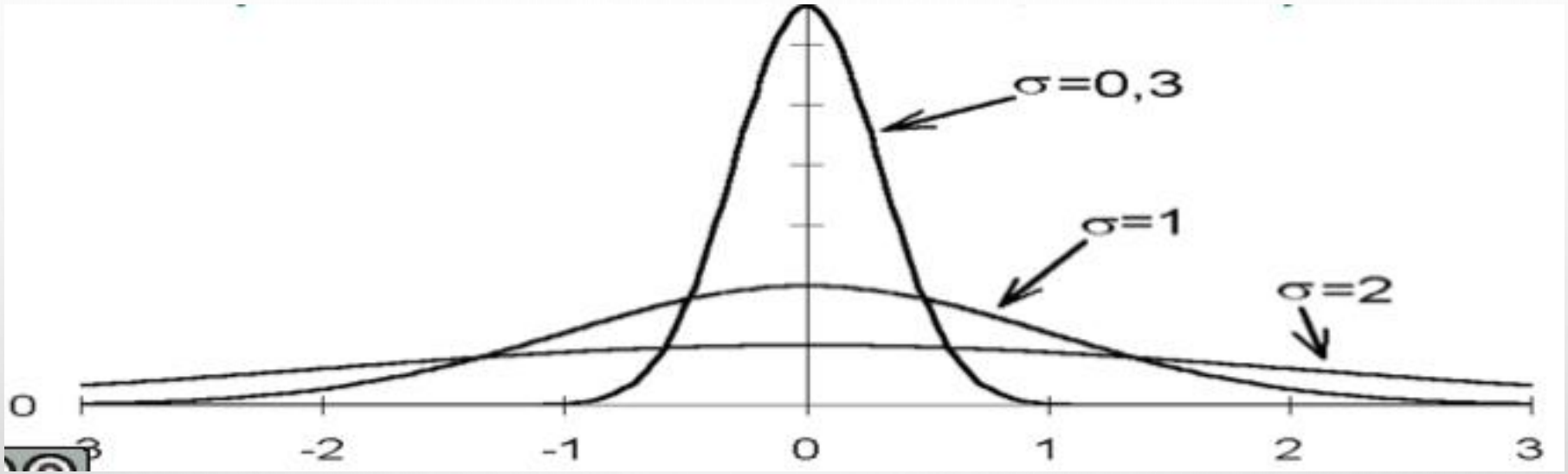


DIFFÉRENTS ASPECTS DE LA COURBE



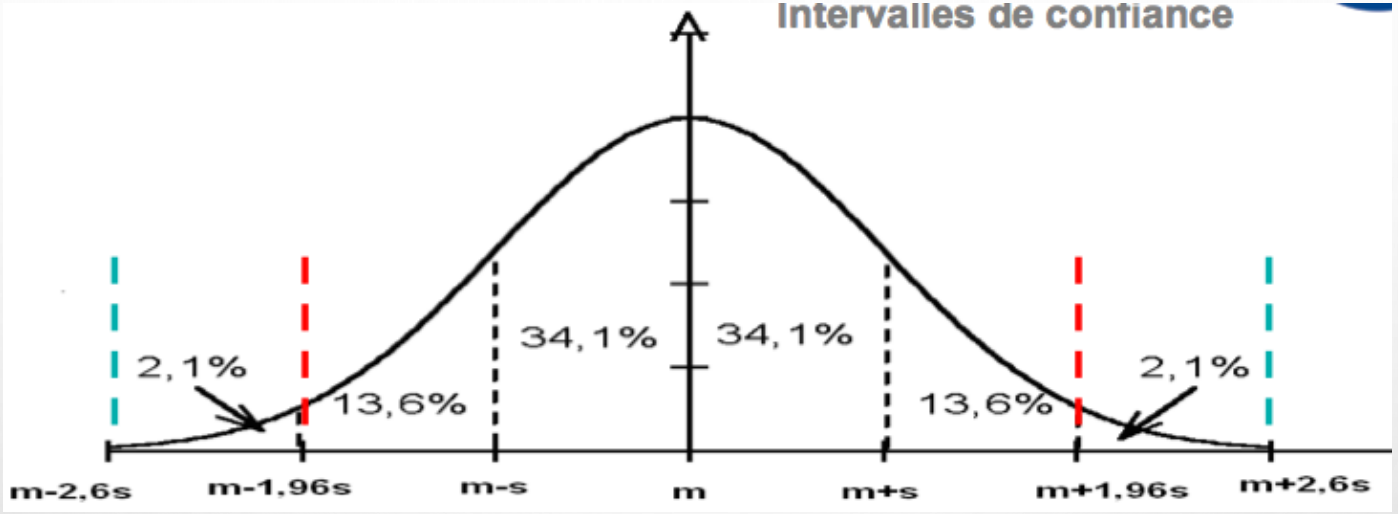
La valeur de la MOYENNE modifie la POSITION de la courbe en cloche :

- La courbe se déplace vers la droite pour les moyennes élevées
- La courbe se déplace vers la gauche pour les moyennes faibles



La valeur de l'ECART TYPE modifie la FORME de la courbe en cloche :

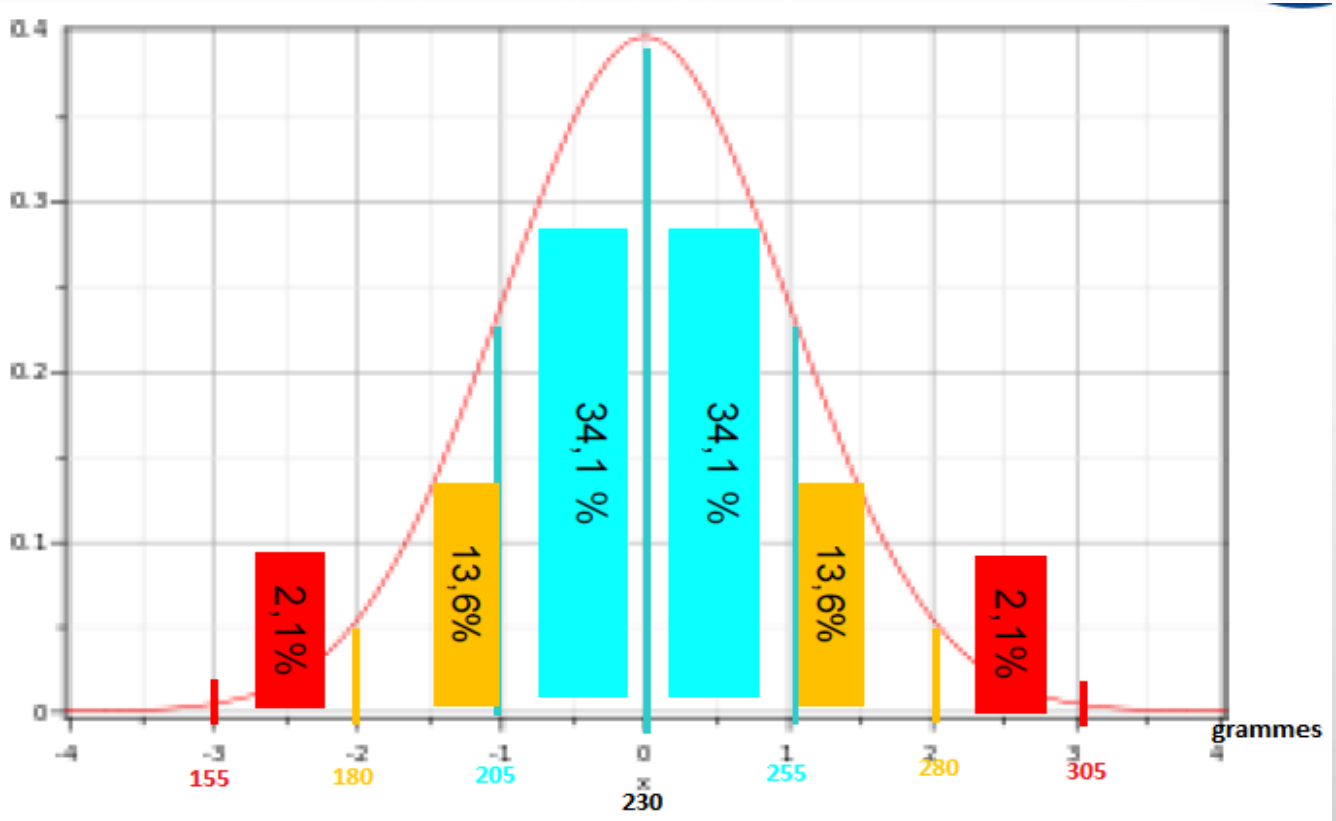
- La courbe s'aplatit pour des valeurs dispersées.
- La courbe se resserre pour des valeurs proches



La répartition des effectifs est proportionnelle à la surface sous la courbe :

- **IC = [m-1s ; m+1s]** contient **68,2%** de la population
- **$IC_{95\%}$ = [m-1,96 s ; m+1,96s]** contient environ **95,4%** de la population
- **$IC_{99\%}$ = [m-2,6 s ; m+2,6s]** contient environ **99,6%** de la population

Dans une boulangerie, le poids moyen des pains est de 230g, avec un ecart type de 25g



% de sachets entre	
205 g et 255 g	68,2%
180 g et 280 g	95%

ESTIMATION DE DONNÉES QUALITATIVES

Dans une population, quel % d'individus présentent un caractère donné ?

- Echantillon représentatif par TAS (n sujets)
- Calcul d'un % qui tend vers la proportion cherchée, mais s'en écarte suivant une variabilité liée au hasard
- Autre échantillon ➔ autre %

p_{obs} => *E*stimateur du pourcentage inconnu (le *% d'individus présentent un caractère donné)*

Estimateur de l'écart type inconnu s => $s = \sqrt{\frac{p0q0}{n}}$

INTERVALLE DE CONFIANCE

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

Soit un groupe de 220 patients, représentatif d'une population rhumatismale (R). On observe 167 cas de rhumatismes inflammatoires.
Quel pourcentage de rhumatismes inflammatoires dans la population R?

1) Estimation ponctuelle p=167/220 = 0,76 soit **76%** **(notre p_{obs})**

2) **Estimation par intervalle**
Nous choisissons le risque α **= 5%,** donc calcul de $IC_{0,95}$
p calculé = 0,76 donc q = 0,24

$$IC_{0,95} = 0,76 \pm 1,96\sqrt{\frac{0,76\times0,24}{220}}$$

$IC_{0,95}$ = [0,70 ; 0,82]

L'estimation par intervalle semble moins précise. Mais si l'on refait ce calcul sur un autre échantillon, cette nouvelle estimation recouvrira la première. Ce sera beaucoup plus rarement le cas avec l'estimation ponctuelle.

La précision est fonction de $s = \sqrt{\frac{p0q0}{n}}$

(donc si n multiplié par 100 ➔ s divisé par 10 ➔ précision augmente facteur 10)

$IC_{0,95} = 0,76 \pm 1,96\sqrt{\frac{0,76\times0,24}{220}}$

$IC_{0,95}$ = [0,70 ; 0,82]

<u>Si n = 22000 (donc x100) :</u>

$IC_{0,95} = 0,76 \pm 1,96\sqrt{\frac{0,76\times0,24}{22000}}$

$IC_{0,95}$ = [0,754 ; 0,766]

INDICE DE PRÉCISION DE L'ESTIMATION DE DONNÉES QUALITATIVES

$$i = \varepsilon.s = \varepsilon.\sqrt{po.qo/n}$$

La précision est d'autant plus grande que i est faible.

$$IC_{(1-\alpha)} = [\ po - \varepsilon.\sqrt{po.qo/n}\ ;\ po + \varepsilon.\sqrt{po.qo/n}\ ] = [\ po - i\ ;\ po + i\ ]$$

La PRECISION et la TAILLE de l'IC varient bien en SENS INVERSE !
La PRECISION augmente dans le MEME SENS que l'effectif de l'échantillon !

Pour une précision donnée : $n = \frac{\varepsilon^2(po.qo)}{i^2}$

NB : i = indice de précision de l'estimation

Précision d'un sondage

900 personnes ont été interrogées sur leur intention de vote à une élection présidentielle qui oppose 2 candidats A et B.

52% (p=0,52) ont déclaré qu'elles **voteraient A**.
Les journaux annoncent que le candidat A arrive en tête.

$IC_{0,95}$ =[0,52 ± 1,96] = **[0,487 ; 0,553]**

Il est faux d'affirmer que A arrive en tête, 52% est dans l'IC autour de 50%

Les 2 candidats peuvent être considérés comme à égalité!

Une non réponse a un sondage constitue toujours un biais

Statistique déductive

Statistique déductive (explicative ou inductive) : Conclusions à partir d'observations et de mesures : **hasard ou autre explication** ?

Dans les statistiques deductives, contrairement aux statistiques descriptives, on essaie, a partir des observations faites, de **tirer des conclusions**, voir s'il existe une **différence entre 2 populations**.

Pour cela, les épidémiologistes utilisent des **tests d'hypotheses**, qui permettent de décider si on garde ou repousse **H0**.

Definition des hypotheses :

H0= hypothese nulle :

- « Il n'y a pas de difference observee entre les deux groupes »
- « Il n'existe pas de lien entre les 2 caracteres etudies, les fluctuations observees sont donc dues au hasard»

H1 = hypothese alternative.

- « Il y a une difference significative entre les deux groupes »
- « Il existe bien un lien entre les 2 caracteres etudies, les fluctuations observees ne sont donc pas dues au hasard. »

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

Il est plus grave de dire qu'un médicament est efficace alors qu'il ne l'est pas (Dire que H1 est vraie alors que H1 est fausse) car c'est donné un produit inefficace voire toxique à un patient, Que dire qu'un médicament est inefficace alors qu'il l'est (Dire que H0 est vraie alors que H0 est fausse) car il n'y a aucun risque pour le patient.

C La notion de risque :

Rappel de statistique descriptive : Lors de l'estimation d'une valeur x par un IC, α represente le **risque d'erreur** dans l'estimation de x, c'est a dire *le risque pour que l'IC ne contienne pas la vraie valeur de x*. Il est generalement fixe a **5%**

En statistique deductive, on a :

1. α ou risque de premiere espece represente : **le risque de rejeter H0 si H0 est vraie**. Ce risque d'erreur est maitrise, c'est a dire qu'il est fixe (le plus souvent a 5%) avant l'application du test statistique.

2. 1 – α represente la **probabilite d'accepter H0 si H0 est vrai**

3. β ou risque de seconde espece represente le **risque d'accepter H0 si H0 est fausse**. Ce risque d'erreur est neglige et peut donc etre assez important.

4. 1- β represente la **puissance du test**. Il s'agit de la **probabilite de rejeter H0 si H0 est fausse.**

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

QCM : Concernant les risques d'erreurs. Donnez les vraies.

A) La probabilité de rejeter H1 si H1 est vraie correspond au risque de première espèce α.
B) La probabilité d'accepter H0, si H0 est fausse correspond au risque de seconde espèce β.
C) La probabilité de rejeter H1, si H1 est vraie correspond au risque de premiere espèce β.
D) La puissance d'un test correspond à 1-α.
E) Aucune de ces réponses n'est correcte.

QCM : Concernant les risques d'erreurs. Donnez les vraies.

A) La probabilité de rejeter H1 si H1 est vraie correspond au risque de première espèce α.

B) La probabilité d'accepter H0, si H0 est fausse correspond au risque de seconde espèce β.

C) La probabilité de rejeter H1, si H1 est vraie correspond au risque de premiere espèce β.

D) La puissance d'un test correspond à $1-\alpha$.

E) Aucune de ces réponses n'est correcte.

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

Il est plus grave de dire qu'un médicament est efficace alors qu'il ne l'est pas (Dire que H1 est vraie alors que H1 est fausse **= Risque α**) car c'est donné un produit inefficace voire toxique à un patient, que dire qu'un médicament est inefficace alors qu'il l'est (Dire que H0 est vraie alors que H0 est fausse **= Risque β**) car il n'y a aucun risque pour le patient.

Etapes d'un test d'hypothese :

- **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 α**
(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.

Etude entre deux caractères qualitatifs: Test de comparaison de pourcentage ou test du χ^2

$$\varepsilon \text{ calculee} = \varepsilon exp = \frac{p_A - p_B}{\sqrt{\frac{p_A q_A}{n_A} + \frac{p_B q_B}{n_B}}}$$ (a ne pas apprendre !!!)

Si ε calculee > ε theorique, alors on accepte H1 (il y a une différence)

Si ε calculée < ε théorique, alors on accepte H0 (il y n'y a pas de difference)

Si échantillon représentatif de la population, on pourra alors extrapoler la conclusion !

QCM time :

Je cherche a déterminer si le fait de travailler la biostat permet ou non d'augmenter ses chances de réussir le concours de paces à Nice. Pour cela, je dispose d'un échantillon de 100 étudiants en paces niçois tiré au sort. Sur 70 etudiants qui travaille la biostat 20 ont réussi le concours. Sur les 30 qui n'ont pas bossé la biostat, 5 ont réussi le concours. On trouve $\varepsilon calc = 2,36$. Que peut on en déduire ?

A) Au risque alpha = 5%, on rejette l'hypothese H0

B) Au risque alpha = 5%, on accepte l'hypothese H0

C) Au risque alpha = 1%, on rejette H0

D) H1 = Sur cette échantillon, les chances de réussir son concours sont differentes en ayant travaillé la biostat qu'en faisant l'impasse.

E) Aucune réponse n'est exacte

QCM time :

Je cherche a déterminer si le fait de travailler la biostat permet ou non d'augmenter ses chances de réussir le concours de paces à Nice. Pour cela, je dispose d'un échantillon de 100 étudiants en paces niçois tiré au sort. Sur 70 etudiants qui travaille la biostat 20 ont réussi le concours. Sur les 30 qui n'ont pas bossé la biostat, 5 ont réussi le concours. On trouve $\varepsilon calc = 2,36$. Que peut on en déduire ?

A) Au risque alpha = 5%, on rejette l'hypothese H0

B) Au risque alpha = 5%, on accepte l'hypothese H0

C) Au risque alpha = 1%, on rejette H0

D) H1 = Sur cette échantillon, les chances de réussir son concours sont différentes en ayant travaillé la biostat qu'en faisant l'impasse

E) Aucune réponse n'est exacte

εcalc = 2,36

On se rappelle que pour $\alpha = 5\%$, εth = 1,96.
Dans ce cas on a εcalc > εth. Donc On accepte H1 et on rejette H0.

Pour $\alpha = 1\%$ on a εth= 2,6. Donc dans ce cas là, εcalc < εth on rejette H1 et on accepte H0.

TAS donc on peut extrapoler la différence a la population, au risque de 5%.

α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0			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=2%, on voit que εth=2,326. On se rappelle que $\varepsilon calc$= 2,36. Donc on a $\varepsilon th < \varepsilon calc$

=> Donc au final, on accepte H1 au risque 2%

Etude entre deux caractères qualitatifs: Test de comparaison de pourcentage ou <u>test du χ^2</u>

La variable Z est ici representee par le χ^2. On comparera :

1. χ^2 theorique est donné par la table du χ^2 en croisant :

- **α (le risque de premiere espece… Generalement 5%)**
- **Le nombre de degre de liberte.**

Pour le test du χ_2, le nombre de ddl vaut : <u>(Tableau !)</u>

$$(nb_{lignes} - 1) \times (nb_{colonnes} - 1)$$

2. $\chi^2_{calcule} = \chi^2_{exp} = \sum \frac{(Oi - Ci)^2}{Ci}$

Si χ^2 calculee > χ^2 theorique, alors on accepte H1 au risque α (il y a une difference)

Si χ^2 calculee < χ^2 theorique, alors on accepte H0 au risque α (il y n'y a pas de difference)

Exemple énoncé **test du χ^2:**

Les telomeres sont des « capuchons » qui protegent l'ADN. Ces bouts retrecissent avec l'age, on pense donc qu'il y a une correlation entre le vieillissement et la taille de ces télomeres. Un nouveau médicament, le telomerase activator permettrait d'augmenter la taille de ces télomeres. Peut etre la future fontaine de jouvence? Pour vérifier cela, on prend 2 groupe de 50 personnes: Un prenant le médicament, l'autre prendra un placebo. On regarde si la longueur des télomeres a augmenté ou non au bout de 3 mois. On obtient $\chi^2_{\text{calculé}} = 3{,}5$

	Augmentation	Non augmentation
Groupe P	6	44
Groupe M	13	37

	Augmentation	Non augmentation
Groupe P	6	44
Groupe M	13	37

Nombre de ddl : 2 lignes, 2 colonnes : (2-1)x(2-1) = 1

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

A α = 5%, χ^2th = 3,841 (1 ddl)
Dans l'énoncé on avait χ^2calc=3,5

Donc?

$\chi^2 th > \chi^2 calc$

$\Rightarrow$ On rejette H1, on accepte H0 donc il n'y a pas de différence significative dans l'échantillon, le médicament est inutile (au risque de 5%).

Peut on extrapoler le résultat à la population Francaise ???

NON, l'énoncé ne nous indique en rien que les groupes on été tiré au sort, et on ne sait presque rien sur l'echantillon…

« **on prend 2 groupe de 50 personnes… »**

Fin !