BIOSTATISTIQUES.03/04/08.

I.Rappels

II.Applications des théories gaussiennes aux échantillons

A.La distribution d'échantillonnage
B.La notion d'erreur type
C.Comparaisons de 2 moyennes.

1-Test paramétrique de student t.

a-Hypothèse nulle et risque alpha.

b-Hypothèse alternative et risque bêta.

c-Test t' de student.
- 2 échantillons appariés-
- 2 échanillons non appariés-

2-Tests non paramétriques

a) 2 échantillons appariés (test de Wilcoxon).
b) 2 échantillons non appariés (test U-Mann-Whitney).
c) Plus de 2 échantillons (test de Kruskall-Wallis).

III.Variable qualitative et test du χ^2 (khi-deux).

A.Tableau de contingence.
B. $\chi^2_{\text{observé}}$ et table du χ^2.

Pour la bonne compréhension de ce cours il est nécessaire d'utiliser des courbes de Gauss (celles des cours précédents par ex).

I. Rappels.

Tout travail statistique, par la force des choses, ne peut porter que sur un nombre limité de valeurs dont l'ensemble constitue l'**échantillon statistique.** Or, ce qui nous intéresse ce n'est pas l'échantillon mais la population d'origine dont il est issu.

Exemple: on étudie le taux de cholestérol chez 30 adultes normaux. Ce qui nous intéresse, ce ne sont pas les 30 personnes de l'échantillon mais l'adulte en général.

Bien entendu on ne peut espérer déterminer avec certitude la valeur véritable du paramètre correspondant de la population d'origine car on ne pourrait pas appréhender l'ensemble des valeurs de la population totale.

Toutefois les méthodes statistiques permettent de déterminer, avec un certain degré de crédibilité, les limites entre lesquelles le paramètre envisagé (de la population d'origine) doit se situer: c'est **l'intervalle de confiance** du paramètre en question.

Usuellement l'intervalle de confiance est **[μ–1,96σ; μ+1,96σ].** (avec μ: la moyenne dans la population et σ: l'écart type). Il est issu de **la loi normale** (Cf.cours précédent).

Graphiquement l'aire sous la courbe située dans cette intervalle de confiance représente **95% de la population** définie comme étant normale (Cf.courbes dans cours précédent).Les 5% restants (considéré comme en dehors de la norme) correspondent à ce qu'on appelle **le risque** α.Ce risque peut être **bilatéral** sur la courbe de Gauss i.e réparti symétriquement des 2 côtés de la moyenne ou **unilatéral** i.e regroupé que d'un coté de la courbe.

II. Application des théories gaussienne aux échantillons.

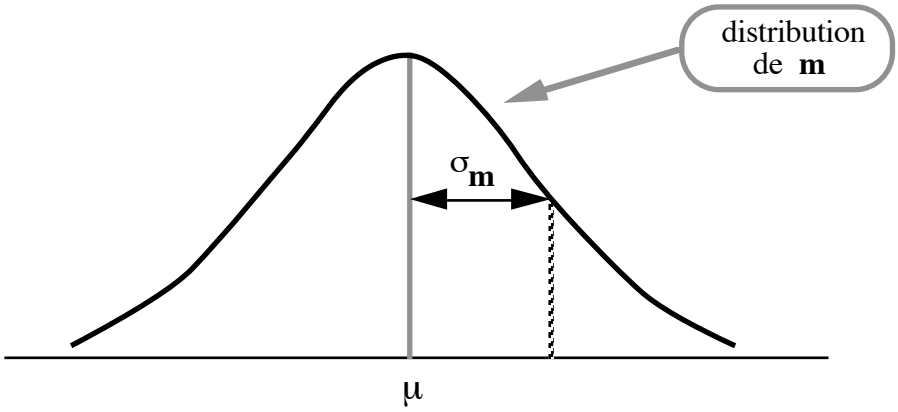
A. La distribution des échantillons.

Un échantillon est défini par **sa taille n** qui est inférieure à celle de la population N dont il est issu (on a n≤N) et par **sa moyenne** $\mathbf{m = \mu_{pop} \pm 1{,}96\sigma_m}$ (autrement dit on considère que la moyenne **m** de l'échantillon appartient à l'intervalle de confiance).
Il existe des échantillons de taille différente et on peut en définir une infinité: c'est ce qu'on appelle **la distribution d'échantillonnage.**

A partir de la moyenne de **la distribution d'échantillonnage**, on va pouvoir définir la notion **d'erreur type** σ_m.

B .La notion d'erreur-type.

L'erreur-type est l'écart-type sur l'ensemble des différentes moyennes calculées sur **la distribution d'échantillonnage**.



Par exemple: on prend un échantillon de 20 étudiants dans une population de 100 et on fait la moyenne d'une variable précise. Puis, on reprend un autre échantillon de 20 étudiants dans cette même population puis on fait de nouveau la moyenne. On peut comprendre que, à priori, ces 2 moyennes ne seront rigoureusement pas identiques sans, pour autant, être très éloignées l'une de l'autre.Cela est du au fait que les échantillons sont différents (dans une certaine mesure): c'est **l'erreur d' échantillonnage**.

L'erreur-type nous permet d'évaluer l'ampleur de ces variations.

Graphiquement, l'erreur-type se lit sur la courbe de Gauss ayant sur l'axe des abscisses les valeurs des moyennes des échantillons (m); elle correspond à l'écart type de cette courbe.Cette courbe ayant pour moyenne la moyenne des moyennes des échantillons i.e la moyenne théorique dans la population.(on comprend que l'erreur type est l'ecart à la moyenne théorique de des moyennes des échantillons).

Mathématiquement l'erreur-type (σ_m) s'exprime en fonction de l'écart-type dans la population (σ_{pop}) et la taille de l'échantillon (n) selon la formule suivante:

$$\sigma_m = \frac{\sigma_{pop}}{\sqrt{n}}$$

or on ne connait pas l'écart type dans la population don on va l'exprimer en fonction de l'écart-type de l'échantillon (s_{ech}), on a alors:

$$\sigma_m = \frac{s_{ech}}{\sqrt{(n-1)}}$$

C. **Comparaison de deux moyennes**.

Quand on veut comparer des moyennes on doit faire face à des problèmes de comparaison.La question générale qui se pose dans ces problèmes c'est de savoir si les échantillons étudiés peuvent être considérés comme différents.En effet, même si les échantillons comparés proviennent d'une même population d'origine, on observe des différences dans leur moyenne;ces différences étant dues **aux fluctuations d'echantillonnage** (les différences entre les échantillons).

Ainsi on veut savoir si les différences observées entre les échantillons ne sont pas simplement dues à ces fluctuations;au quel cas **elles ne devraient pas être prises en compte**.Si, au contraire, les différences observées sont trop importantes pour être mises sur le compte des fluctuations d'echantillonage, on dira qu'elles sont **significatives**. (Cf.plus tard)

1- Test paramétrique de Student t

Le t**est paramétrique de Student** repose sur des **comparaisons de moyennes** (entre la moyenne d'un échantillon et une moyenne théorique ou bien entre les moyennes de 2 échantillons).Le test de Student permet ainsi de comparer des échantillons indépendants et/ou appariés.
Pour appliquer ce test il faut que la distribution de l'échantillon **suive la loi normale de Gauss**.Ainsi il faut notamment que sa **taille soit supérieure ou égale à 30** ($n \geq 30$).De plus ce test n'est applicable que pour une **variable quantitative.**
Mathématiquement, sa formule générale est la suivante:

$$t = \frac{\left|\mu_{theo} - m_{ech}\right|}{\frac{s}{\sqrt{n-1}}}$$

(formule à connaître par coeur)

Avec m_{ech} (moyenne de l'échantillon), μ_{theo} (moyenne théorique).
A noter que si on compare deux échantillons cette moyenne sera remplacée dans la formule par celle d'un échantillon.Le dénominateur correspond à l'erreur type vue précédemment.

a- Hypothèse nulle H_0 et risque α (ou de première espèce)

L'hypothèse nulle Ho est la suivante:
On considère deux échantillons A et B de moyennes respectives m_A et m_B et on pose $m_A - m_B = 0$. Autrement dit on considère qu'il n'existe pas de différence significative entre les deux échantillons (mise à part les différences dues aux fluctuations d'echantillonnage).

On rappelle que **l'intervalle de confiance** usuellement utilisé est $[\mu - 1{,}96\,\sigma\ ;\ \mu + 1{,}96\,\sigma]$. De plus, d'après les propriétés de la **distribution normale**,pour deux échantillons provenant d'une même population d'origine(comme on en a fait l'hypothèse au préalable) une **différence** $d \geq 1{,}96\sigma$ (donc qui sort de l'intervalle de confiance) ne sera due aux **fluctuations d'échantillonage** que dans **moins de 5% des cas**.
Ainsi, comme on considère que les deux échantillons proviennent d'une même population d'origine,on peut dire qu'**il n'y a que moins de 5% de chance qu'il existe une différence** $d \geq 1{,}96\sigma$.Ainsi si on trouve une différence $d \geq 1{,}96\sigma$ entre les deux échantillons testés, on préfère **considérer** que ces **2 échantillons proviennent de**

populations d'origine différentes (DONC que les différences sont significatives) plutôt que de mettre cette différence sur le compte des fluctuations d'echantillonnage.
En agissant ainsi, on prend tout de même le **risque de rejeter H_0 alors que H_0 est vraie**(i.e que cette différence $d \geq 1{,}96\sigma$ soit due aux fluctuations d'échantillonnage): **c'est le risque α**. (usuellement il est inférieur ou égal à 5% ou encore à 0,05).

Remarque: si on élargit l'intervalle de confiance par exemple à 99% ou encore **$[\mu - 2{,}58\,\sigma\ ;\ \mu + 2{,}58\,\sigma]$**. Une différence $d \geq 2{,}58\sigma$ entre deux échantillons provenant d'une même population d'origine, ne sera due aux fluctuations d'echantillonnage que dans moins de 1% des cas.Le risque α sera alors≤1%.Le test sera alors plus précis.
Le choix du risque □ est fait par l'expérimentateur,selon le matériel utilisé,pour obtenir des informations utilisables.

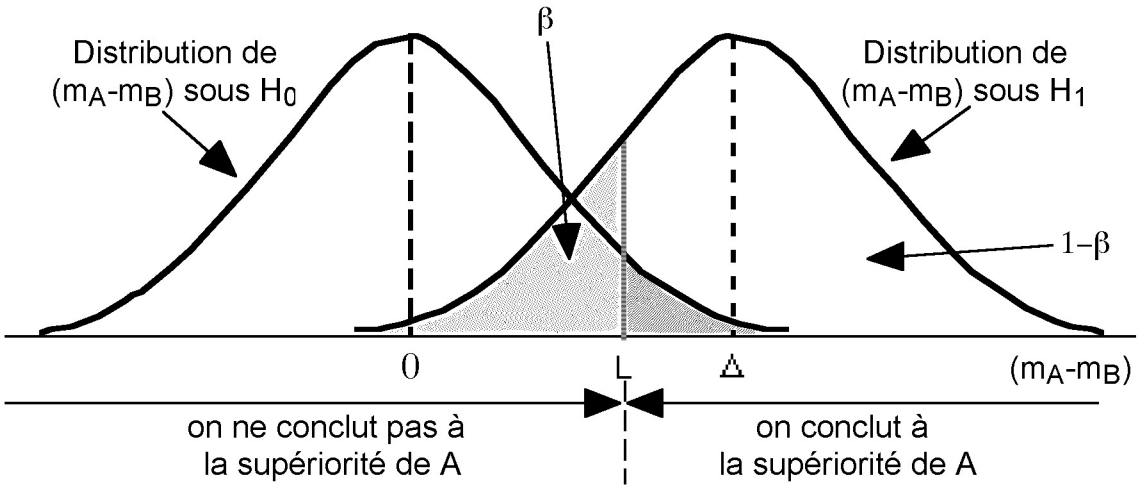
b- Hypothèse alternative H_1 et risque β (ou de deuxième espèce).

L'hypothèse alternative H_1 est la suivante :

On considère toujours les deux mêmes échantillons A et B mais cette fois-ci on pose :
$m_A - m_B \neq 0$. Autrement dit on considère que les deux échantillons proviennent de populations d'origines différentes.

Une fois de plus on définit un risque: **le risque β**. Le risque β est la probabilité de **ne pas rejeter H_0 alors que H_1 est vraie**. Autrement dit, le risque c'est de considérer qu'il n'existe pas de différences significatives entre les échantillons alors qu'il en existe une.Le risque β est usuellement égal à 10% (ou d'autres valeurs > 5%).

A partir de ce risque β, on définit l**a puissance du test: 1 - β.**



Résumé des hypothèses et des risques, puissance:

		Décision	
		rejet de H_0	**non rejet de H_0**
Réalité	**H_0 est vraie**	Erreur de 1ère espèce risque α	$1 - \alpha$
	H_1 est vraie	Puissance $1 - \beta$	Erreur de 2ème espèce risque β

Remarque:plus on diminue le risque α, plus on augmente le risque β.
Ce qu'il faut retenir:
- **si p<0,05, la différence est significative.**
- **si p≥0,05, la différence n'est pas significative.**

pour t: **-si t<1,96, la différence n'est pas significative.**
-si t≥1,96, la différence est significative.

c- Test paramétrique t' de Student

La formule du test (vue plus haut) varie selon le type d'échantillon utilisé.

- **2 échantillons appariés** -

Ici, le test concerne en **fait un échantillon étudié à deux temps différents.**
Par exemple: on dose la glycémie sur un échantillon de 30 personnes à t_o, puis on la redose sur ce même échantillon à t_1 = 6mois.

Ainsi, l'échantillon a pour témoin lui même mais à un temps différents.
Pour ce type d'échantillons la formule du test se formule:

$$t' = \frac{|m_0 - m_1|}{\sqrt{\frac{s_0^2 + s_1^2}{n - 1}}}$$

(formule à connaître par coeur)

- **2 échantillons non appariés** -

Ici, le test concerne **2 échantillons différents**.
Par exemple: on dose la glycémie sur un échantillon **1** (n_1,m_1,s_1)
et sur un échantillon **2** (n_2,m_2,s_2)
pour les comparer, on utilise la formule suivante:

$$t' = \frac{|m_1 - m_2|}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}}$$

(formule à connaître par coeur)

2- Tests non paramétriques.

Si la variable **ne suit pas la loi normale** (autrement dit que la représentation graphique n'est pas une courbe de Gauss ou encore **m ≠ Médiane**), le **test de Student** n'est **pas applicable**.
Les comparaisons entre les échantillons seront faites grâce à des logiciels informatiques.
Pour appréhender ce cas de figure le statisticien anglais **Wilcoxon** a décidé d'utiliser des rangs plutôt que des variables
Par exemple :: 2 étudiants en PCEM1, Jérémie et Maxime, reçoivent leur note du concours.
Jérémie a les notes suivantes: 8, 10, 12, 14,16.
Maxime a les notes suivantes: 7, 9,11, 13, 17
A chaque note on va attribuer un rang :

7 —>*1*	**12** —>*6*
8 —>*2*	**13** —>*7*
9 —>*3*	**14** —>*8*
10 —>*5*	**16** —>*9*
11 —>*5*	**17** —>*10*

Puis, on va mélanger les notes pour avoir un échantillon plus large.On considère les rangs mais pas les notes.Puis on utilise un logiciel pour comparer.

Avec ces rangs de Wilcoxon on peut avoir 3 situations.

a- **2 échantillons appariés.**

On l'appelle le **test de Wilcoxon**.Le test est réalisé par un logiciel informatique.
Si p < 0,05: la différence est significative.

b- **2 échantillons non appariés**.

On l'appelle **test U-Mann-Whitney**. Le test est réalisé par un logiciel informatique
Si p < 0,05: la différence est significative.

c- **Plus de 2 échantillons**.

On l'appelle **le test de Kruskall-Wallis** .Le test est réalisé par un logiciel informatique.
Si p < 0,05: la différence est significative.

Remarque: **les tests paramétriques sont plus fiables (on dit qu'ils sont plus robustes) que les tests non paramétrique**s car ils reposent sur **la loi normale**.
Par exemple si ces deux types de test donnent des résultats contradictoires, on se fie au test paramétrique.

III- Variables qualitatives, test du χ^2 (lire khi-deux).

Pour les variables qualitatives, on utilise le test du χ^2.

1- Tableau de contingence (ou tableau des observations).

Par exemple pour les colonnes la couleur des cheveux et pour les lignes la couleur des yeux.

Nous allons utiliser le test du χ^2 pour confronter plusieurs distributions observées entre elles, c'est à dire pour tester l'hypothèse **H_0 qu'elles proviennent d'une même population**. La difficulté qui demeure toujours la même réside dans le fait que la population n'est pas connue à priori et qu'il va falloir étudier sa composition.

Couleur des cheveux / Couleur des yeux	Blonds	Bruns	Noirs	Roux	TOTAL
Bleus	25	9	3	7	44
Gris ou verts	13	17	10	7	47
Marrons	7	13	8	5	33
TOTAL	45	39	21	19	**124**

Ce tableau indique la répartition de 124 sujets classés d'après la couleur de leurs yeux et la couleur de leurs cheveux.
On se demande si les 3 échantillons de sujets "yeux bleus", "yeux gris ou verts", "yeux marrons" sont comparables quant à la distribution des couleurs de cheveux.
En terme de population, on va se demande si les 3 échantillons proviennent de la même population à 4 couleurs de cheveux (blond, brun, noir, roux).

Dans un premier temps on va estimer ce que serait la composition de la population à 4 couleurs de cheveux. La composition la plus probable déduite de l'ensemble des mesures est

Blonds	Bruns	Noirs	Roux	TOTAL
$p_1 = \frac{45}{124}$	$p_2 = \frac{39}{124}$	$p_3 = \frac{21}{124}$	$p_4 = \frac{19}{124}$	100%

Il s'agit donc de décider si les 3 échantillons observés proviennent de la population définie par P_1 , P_2 , P_3 , P_4 .

Les effectifs calculés du 1er échantillon de 44 sujets sont ceux qui correspondent à la composition de la population soit

Blonds	Bruns	Noirs	Roux	TOTAL
$44\ p_1 = 44 \times \frac{45}{124}$	$44\ p_2 = 44 \times \frac{39}{124}$	$44\ p_3 = 44 \times \frac{21}{124}$	$44\ p_4 = 44 \times \frac{19}{124}$	**44**

De même on calcule les effectifs théoriques du 2ème et 3ème échantillon.

On obtient alors le **tableau dit "théorique"**

Couleur des cheveux / Couleur des yeux	Blonds	Bruns	Noirs	Roux	TOTAL
Bleus	16,0	13,8	7,5	6,7	44
Gris ou verts	17,1	14,8	7,9	7,2	47
Marrons	11,9	10,4	5,6	5,1	33
TOTAL	45	39	21	19	**124**

Dans un deuxième temps on va chercher si les écarts du 1er tableau par rapport à la distribution théorique du 2ème tableau sont compatibles ou non avec les fluctuations d'échantillonnage.

On va calculer le χ en prenant la somme des

$$\frac{\left(n_{obs} - n_{theor}\right)^2}{n_{theor}}$$

de chacune des 4 x 3 = 12 cases définis par les 4 couleurs de cheveux et des 3 couleurs d'yeux, on trouve $\chi^2_{calc} = 15$.

$$\chi^2_{calc} = \sum \frac{\left(n_{obs} - n_{theor}\right)^2}{n_{theor}}$$

Pour savoir si ce $\chi^2_{calc} = 15$ est trop grand ou non, on doit regarder dans la table du χ^2_{theor} avec un nombre de degrés de liberté convenable. On montre, de façon générale,

que pour un tableau L lignes et C colonnes. Ce nombre de degrés de liberté est

ν = (L-1) (C-1)

Ici on a donc ici ν = (L-1) (C-1) = 6

Pour ν = 6 et un risque α de 5 % la table indique χ^2_{theor} **= 12,592**. La valeur trouvée étant plus grande on doit rejeter l'hypothèse H_0 selon laquelle les trois groupes "yeux bleus", "yeux verts ou gris", "yeux marrons" proviendraient de populations identiques quant à la répartition des couleurs de cheveux ⇒⇒⇒⇒ ces répartitions diffèrent significativement.

Méthode générale

Nous disposons au départ d'un **tableau de contingence** à L lignes et C colonnes.
Nous devons calculer les effectifs théoriques de toutes les cases : l'effectif calculé d'une case est le produit du total de sa ligne par le total de sa colonne, divisé par le total général.

Nous avons ensuite effectué la somme des

$$\frac{\left(n_{obs} - n_{theor}\right)^2}{n_{theor}}$$

de chaque case.

Enfin nous avons cherché la probabilité correspondant à ce χ^2_{calc} dans la table du χ^2_{theor} pour **ν = (L-1)(C-1)**.

On peut considérer que le test effectué vise à rechercher s'il existe ou non une liaison entre 2 caractères qualitatifs. Ce test est qualifié de **test d'indépendance** (entre 2 variables qualitatives).

Voilà un extrait de cette table du χ^2.

Degré de liberté ν	pour p= 0,05	pour p= 0,1	pour p=0,001
1	3,84	6,64	10,83
2	5,99	9,21	13,82
3	7,82	11,35	16,27
4	9,49	13,28	18,47
5	11,07	15,09	20,52
6	12,59	16,81	22,46
7	14,07	18,48	24,32
8	15,51	20,09	26,13