Université d'Aix-Marseille 2012–2013

Mathématiques pour la Biologie II

Résumé de cours.

Tests sur les moyennes, variances et proportions.

1 Quatre familles de lois à connaître.

Elles sont classées par ordre d'importance. Il faut savoir un peu de leurs propriétés, et surtout dans quels tests elles interviennent.

1.1 Les lois Gaussiennes $\mathcal{N}(\mu, \sigma^2)$.

- Distributions très répandues et utiles car elles possèdent un caractère universel. Cela est dû au Théorème Centrale Limite.
- Une famille à deux paramètres : la moyenne μ, la variance σ^2.
- La densité de la Gaussienne centrée réduite $\mathcal{N}(0,1)$ est $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Si $x \sim \mathcal{N}(0,1)$, et $a < b$ sont deux réels, alors
$$\mathbb{P}\big(X \in [a,b]\big) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}x^2}\, dx.$$
- Si $X \sim \mathcal{N}(\mu, \sigma)$, alors $\dfrac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.
- Si $X_1, \ldots, X_n$ sont n v.a. indépendantes de loi $\mathcal{N}(\mu, \sigma)$, alors $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ suit la loi $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

1.2 Les lois du χ^2.

- Distributions très utiles en statistiques pour leur lien avec les lois Gaussiennes et leurs propriétés.
- C'est une famille de lois à un paramètre : le **nombre de degrés de liberté** (**d.d.l.** par la suite).
- **Définition :** La loi du χ^2_ν, à ν d.d.l, est celle de la variable
$$Y = X_1^2 + \ldots + X_\nu^2, \qquad \text{avec des } X_i \text{ indépendantes toutes de loi } \mathcal{N}(0,1).$$
- La moyenne du χ^2_ν est exactement ν. Sa variance vaut 2ν. Donc si ne nombre de d.d.l. ν est assez grand, alors on peut approcher (TCL) par la loi normale $\mathcal{N}(\nu, 2\nu)$.
- Une variable qui suit une loi χ^2_ν ne prend que des valeurs positives.
- La somme de deux variables indépendantes suivant une loi du χ^2 suit encore une loi du χ^2. Les nombres de degrés de liberté s'additionnent.

1.3 Les lois de Student.

- Distributions utilisées en statistiques pour estimer, tester, **comparer des moyennes quand on ne connaît pas la variance théorique**.
- C'est une famille de lois à un paramètre : le nombre de degrés de liberté.

- **Définition :** La loi du Student à ν d.d.l, notée $\mathcal{S}_\nu$ est celle de la variable

$$Y = \frac{X_1}{\sqrt{\frac{1}{\nu}X_2}}, \qquad \text{lorsque } X_1 \sim \mathcal{N}(0,1),\ X_2 \sim \chi^2_\nu,\ X_1 \text{ et } X_2 \text{ sont indépendantes.}$$

- Toutes les lois de student sont centrées (de moyenne nulle) et symétriques (leur représentation grahique est symétrique). Leur variance vaut $\frac{\nu}{\nu-2}$, si $\nu \geq 3$.
- Pour $\nu \geq 30$ (ou mieux 60), la loi $\mathcal{S}_\nu$ est assez bien approchée par la loi $\mathcal{N}(0,1)$.

1.4 Les lois de Fisher(-Snedecor).

- Distributions utlisées pour tester l'égalité des variances de deux lois.
- Une famille de loi à deux paramètres.
- **Définition :** La loi de Fisher $\mathcal{F}_{(\nu_1,\nu_2)}$ de paramètres (ν_1,ν_2) est la loi de

$$Y = \frac{\frac{1}{\nu_1}X_1}{\frac{1}{\nu_2}X_2}, \qquad \text{lorsque } X_2 \sim \chi^2_{\nu_1},\ X_2 \sim \chi^2_{\nu_2},\ X_1 \text{ et } X_2 \text{ sont indépendantes.}$$

- la moyenne de $\mathcal{F}_{(\nu_1,\nu_2)}$ est $\dfrac{\nu_2}{\nu_2-2}$ pour $\nu_2 \geq 3$. Une valeur proche de 1 lorsque ν_2 est grand.
- Une variable qui suit la loi $\mathcal{F}_{(\nu_1,\nu_2)}$ ne prend que des valeurs positives.

Fractile : Pour un nombre $\gamma \in [0,1]$, et une variable aléatoire Z qui suit une loi donnée (dans les cas qui nous intéressent $\mathcal{N}(0,1)$, Student, χ^2 ou Fisher), on définit le fractile d'ordre γ de la loi comme le nombre réel z_γ tel que la probabilité d'avoir $Z \in (-\infty, z_\gamma)$ vaut exactement γ.

$$\mathbb{P}\left(Z \leq z_\gamma\right) = \gamma.$$

Les fractiles de ces différentes lois sont donnés par des tables ou calculés par ordinateur. Pour les tables de ces 4 lois, on peut consulter par exemple :

http://www.itl.nist.gov/div898/handbook/eda/section3/eda367.htm

2 Deux estimateurs importants.

En statistique, on cherche à obtenir des informations sur une grandeur (taille, poids, proportion quelconque...) sur une grande population (de taille N). Comme il est impossible de connaître les grandeurs associées à chaque individu, on s'intéresse à la distribution (ou répartition) de cette grandeur. Sauf mention contraire explicite, dans la suite on supposera que sa distribution de cette valeur est Gaussiennce de moyenne μ et écart-type σ : $\mathcal{N}(\mu,\sigma^2)$. Le Théorème Central de la Limite assure que cette hypothèse est acceptable dans de nombreuses situations.

Lors d'une expérience sur un échantillon de taille n issu de la population, on mesure n grandeurs $(x_1,\ldots,x_n)$. On peut alors calculer une moyenne et un écart-type empirique modifié.

$$\text{la moyenne empirique : } \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \text{l'écart-type empirique modifié : } s_n^2 = \frac{1}{\mathbf{n-1}}\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2.$$

Ces valeurs empiriques obtenues sur un échantillon sont-elles pertinentes, ou dépendent-elles de l'échantillon ? Aurait-on obtenu des valeurs proches ou différentes si on avait choisi un autre échantillon ?
Pour répondre à cette délicate question, il faut faire une expérience de pensée, et considérer un échantillon aléatoire, dont les résultats sont n variables aléatoires X_i supposés indépendantes (en général), toutes de même loi inconnue : ici $\mathcal{N}(\mu, \sigma^2)$ avec μ et σ inconnus.
Pour cet échantillon aléatoire, les moyennes empiriques et l'écart-type empiirique modifié $\bar{X}_n$ et S_n sont deux variables aléatoires et :

- la moyenne empirique $\bar{X}_n$ suit la loi Gaussienne $\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$,
- la variance empirique modifiée : S_n^2 suit une loi du χ^2 à $\nu = n-1$ **degrés de liberté**.

Cas des proportions : On peut se baser sur ce qui précède car une proportion est une moyenne de variables qui ne prennent que deux valeurs : 0 ou 1.

3 Tests d'hypothèses.

Pour un test d'hypothèse classique, il existe deux possibilité :

- on pécise un risque seuil α petit : 10%, $5\%, 1\%$:..
- on ne précise pas de risque seuil α mais on calcule la **valeur p**.

Que représente le risque seuil α (et la valeur p) ? Contrairement à une croyance répandue, le risque seuil $\alpha = 5\%$ ne veut pas dire que l'hypothèse à 95% de chances d'être vrai quand le test est concluant. Il suffit de regarder la construction d'un test ci dessous pour s'en convaincre : le risque ne devient une probabilité qu'une fois l'hypothèse acceptée. C'est la même chose pour la valeur p.
Par contre, on peut dire la chose suivante : sur 1000 tests indépendants au risque seuil de 5% dont les hypothèses sont vraies, on va rejeter (H_0) (donc à tort) dans environ 5% des cas, donc environ 50 fois ici d'après la loi des grand nombres. Toutefois, il faut prendre ce calcul avec précaution, car on ne peut savoir à l'avance si les hypothèses d'un test sont vraies.
Ceci est un vrai problème pour les très nombreuses publications scientifiques en médecine ou biologie, qui comporte presque toujours des tests statistiques. Il en existe forcément qui ont rejeté des hypothèses à tort. Pour cette raison, on préfère souvent donner la valeur p, qui permet de mesurer à quel point on peut se fier au résultat.

3.1 Principe général.

1. On choisit une hypothèse (H_0) que l'on va tester. Son contraire est noté (H_1).
2. On donne la formule d'une **variable de décision** Z, aussi appelée **statistique du test**, qui dépend des résultats de l'expérience. Dans l'expérience de pensée avec échantillon aléatoire, cette variable est aléatoire. Et **si l'hypothèse** (H_0) **est vérifiée**, elle a une loi bien connue : loi normale centrée réduite, loi du χ^2, loi de Student ou loi de Fisher. On connait donc ses fractiles (tables ou ordinateur).
3. On cherche la **zone de rejet** associée qui contiennent la proportion α des valeurs extrèmes, pour la loi considerée. Un dessin est souvent utile.
   - Pour un test bilatéral, les valeurs extrèmes à rejeter sont situées en-dessous du fractile $z_{\alpha/2}$ et au-dessus de $z_{1-\alpha/2}$.
   $$R =]-\infty, z_{\alpha/2}] \cup [z_{1-\alpha/2}, +\infty[.$$
   Par exemple pour un seuil de 5%, on utilisera $z_{0,025}$ et $z_{0,975}$.
   - Pour un test unilatéral, on utilisera soit z_α ou $z_{1-\alpha}$, mais pas les deux. Le choix dépend de l'inégalité qu'on teste.
   $$R =]-\infty, z_\alpha], \quad \text{ou} \quad R = [z_{1-\alpha}, +\infty[.$$

4. On calcule la valeur particulière de la variable de décision z_{emp} obtenue en utilisant les données de l'expérience et aussi l'hypothèse (H_0) si besoin.
   - Si cette valeur fait partie de la zone de rejet, **on rejette l'hypothèse au risque seuil en question**.
   - Si cette valeur est hors de la zone de rejet, on accepte l'hypothèse au risque seuil en question, ou mieux **on ne rejette pas l'hypothèse au risque seuil en question**. L'hypothèse n'a pas été invalidée par ce test, ce qui ne veux pas dire qu'elle pourra être invalidée par une expérience ultérieure.
5. Pour finir on peut calculer la **valeur** p, qui est la proportion de valeur plus extrêmes (dans l'expérience de pensée) que celle obtenue sur l'échantillon. Par exemple, si la variable de décision $Z \sim \mathcal{N}(0,1)$, alors
$$\text{valeur } p = \mathbb{P}\big(|Z| \geq |z_{emp}|\big).$$
Le calcul de la valeur p est toujours recommandé : il donne beaucoup plus d'information que le simple fait de passer ou pas le test avec un risque seuil fixé.

Test bilatéral ou unilatéral ? Comment choisir entre les deux ? Cela dépend surtout de la question que l'on se pose. Par exemple, si l'expérience cherche à démontrer qu'un médicament est plus efficace qu'un autre, il faut mieux faire un test unilatéral : on cherche en effet juste à savoir s'il y a plus de guérisons si on utilise le premier médicament. Par contre, si l'expérience est faite pour savoir s'il y a approximativement autant de carpes que de goujons dans un lac, il faut faire un test bilatéral : on veut savoir si les deux espèces ont des populations de taille à peu près égale.

Risque α et risque β ? Le risque que l'on considère ici est le risque de rejeter à tort l'hypothèse (H_0) quand elle est juste. Il est traditionnellement appelé risque α. Pour le diminuer, il suffit simplement de diminuer le risque-seuil α. Le risque β est le risque d'accepter à tort (H_0). Il n'est pas pris en considération dans les tests que vous apprendrez cette année : il se peut donc que nos test donne des résultat positifs à tort. C'est pour cela qu'on préfère l'expression du type "on ne rejette pas l'hypothèse".

4 Les différents tests paramétriques à connnaître.

Dans cette partie, on donne pour les différents tests, les variables de décisions Z à construire et les lois associées.

Notation (utile pour la suite) : Dans le cas où il y a deux grandeurs à comparer, la seconde est notée Y. On enlève les indices n sur les moyennes empiriques $\bar{X}, \bar{Y}$ et les écarts-types empiriques modifiés S_x, S_y. Les moyennes et écats-types sur les populations sont également affublés d'un indice : $\mu_x, \sigma_x, \mu_y, \sigma_y$.

4.1 Conformité d'une proportion à une valeur p_0 connue.

Dans ce cas, la distribution de X est une loi de Bernouilli $\mathcal{B}(p)$.

- (H_0) : $p = p_0$
- Conditions d'application : $n \geq 30$, $np_0 \geq 5$, $n(1-p_0) \geq 5$,
- $Z = \sqrt{n}\frac{\bar{P}_n - p}{\sqrt{p(1-p)}} \sim \mathcal{N}(0,1)$.

4.2 Comformité d'une moyenne à une valeur μ_0 connue.

- $(H_0) : \mu = \mu_0$,

- $Z = \sqrt{n}\dfrac{\bar{X}_n - \mu_0}{S}_n \sim \mathcal{S}_{n-1}$, (Student à $n-1$ d.d.l.)

4.3 Comformité d'un écart-type à une valeur σ_0 connue.

- $(H_0) :\ \sigma = \sigma_0$
- $Z = (n-1)\dfrac{S^2}{\sigma_0^2} \sim \chi^2_{n-1}$, le χ^2 à $(n-1)$ d.d.l.

4.4 Comparaison de deux proportions, échantillons indépendants.

Dans ce cas, la distribution de X et de Y sont des lois de Bernouilli $\mathcal{B}(p_x)$ et $\mathcal{B}(p_y)$.

- $(H_0) :\ p_x = p_y$
- Conditions d'application : $n_x \geq 30$, $n_x\bar{P}_x \geq 5$, $n_x(1-\bar{P}_x) \geq 5$, et les mêmes inégalités avec y
- On pose $\bar{P} = \dfrac{n_x\bar{P}_x + n_y\bar{P}_y}{n_x + n_y}$
- $Z = \sqrt{\dfrac{n_x n_y}{n_x + n_y}}\dfrac{\bar{P}_x - \bar{P}_y}{\sqrt{\bar{P}(1-\bar{P})}} \sim \mathcal{N}(0,1)$.

On peut aussi utiliser le test du χ^2 d'adéquation entre les deux proportions (voir plus loin).

4.5 Comparaison de deux moyennes, pour échantillons indépendants.

On suppose que les écarts-type théoriques des deux populations sont égaux $\sigma_x = \sigma_y$ (sinon c'est plus dur).

- $(H_0) : \mu_x = \mu_y$, $\mu_x \leq \mu_y$, $\mu_x \geq \mu_y + 1$...
- On définit l'écart-type empirique modifié de la réunion des deux échantillons $S = \dfrac{(n_x-1)S_x + (n_y-1)S_y}{n_x + n_y - 2}$.
- $Z = \sqrt{\dfrac{n_x n_y}{n_x + n_y}}\dfrac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{S} \sim \mathcal{S}_{n_x+n_y-2}$.
- Si $n_x + n_y - 2 \geq 30$ (ou mieux 60), on peut remplacer $\mathcal{S}_{n_x+n_y-2}$ par la loi $\mathcal{N}(0,1)$.

Que faire pour un test unilatéral $(H_0) :\ \mu_x \leq \mu_y$?
- Si $\bar{x}_n \leq \bar{y}_n$, on accepte (H_0) sans faire de calculs, puisque les résultats de l'expérience vont dans le sens de l'hypothèse.
- Si $\bar{x}_n \geq \bar{y}_n$, on choisit parmi (H_0) le cas le plus favorable, c'est-à-dire le cas $\mu_x = \mu_y$. On poursuit ensuite comme ci-dessus, mais avec une zone de rejet de la forme $\mathcal{R} =]z_{1-\alpha}, +\infty[$.

4.6 Comparaison de moyennes pour des échantillons appariés.

On utilise ce test quand on veut comparer les moyennes de deux grandeurs mesurées sur un même échantillon (souvent à des temps distincts, par exemple avant et après une expérience). Attention, il faut connaître les données complètes de l'expérience, c-à-d les résultats x_i et y_i pour chaque individu. Il suffit alors
- de calculer les écarts $d_i = x_i - y_i$ pour chaque individus,
- de calculer ensuite la moyenne empirique et l'écart-type empirique modifié des $(d_i)_{1\leq i\leq n}$.
- d'appliquer aux écarts d_i le test de comparaison de moyenne à une valeur théorique donnée, souvent 0.

4.7 Comparaison de deux écarts-type, échantillons indépendants.

- $(H_0) :\ \sigma_x = \sigma_y$

- $Z = \dfrac{S_x'^2}{S_y'^2} \sim \mathcal{F}_{(n_x-1,n_y-1)}$, la loi de Fisher-Snedecor de paramètre $(n_x - 1, n_y - 1)$.

5 Tests non paramétriques.

On parle de tests non paramétriques lorsque la variable qui nous intéresse n'est plus une grandeur qui peut varier continûment comme une taille, une proportion,... mais une grandeur discrète : par exemple, un choix parmi plusieurs espèces de poissons. Attention, on peut aussi discrétiser une variable continue : par exemple une taille peut-être remmplacée par plusieurs catégories "taille $\leq$ 150 cm, taille $\in [150cm, 180cm]$, taille $\geq 180cm$".

5.1 Test du χ^2 d'adéquation.

Les données de l'expérience sont classées en k catégories, numérotées de 1 à k. On veut comparer cette répartition à une répartition théorique connue : uniforme, loi de Mendel, distribution binomiale, ou parfois une autre répartition expérimentale.... On suit alors les étapes suivantes :

- On remplit le tableau ci-dessous. La première ligne contient les données de l'expérience pour l'échantillon de taille n. La seconde contient les effectifs attendus en moyenne pour un échantillon de même taille n, qui suivrait la loi que l'on cherche à tester.

Catégorie	Cat. 1	Cat. 2	...	Cat k.	Total
Expérience	n_1	n_2	...	n_k	n
Théorie	n_1^{th}	n_2^{th}	...	n_k^{th}	n

- (H_0) : La variable étudiée suit la loi en question (qui sert pour construire la ligne des effectifs théoriques).
- Conditions d'application : $n \geq 30$ et tous les effectifs théoriques sont supérieurs à 5.
- La variable de décision $Z = \displaystyle\sum_{i=1}^{k} \frac{(N_i - n_i^{th})^2}{n_i^{th}}$ suit la loi du χ^2 à $(k-1)$ d.d.l.

5.2 Test du χ^2 d'indépendance.

On utilise ce test pour déterminer si des variables sont indépendantes on non. Au cours d'une expérience, on étudie une variable A qui peut prendre k valeurs distinctes $(A_1, \ldots, A_k)$ **et** une variable B qui prend l valeur distinctes $(B_1, \ldots, B_l)$. Un exemple typique est A prend la valeur "oui" on "non" suivant qu'on a pris traitement ou pas et B donne le résultat du traitement "toujours malade", "guéri" ou "mort"... La procédure est la suivante.

- On remplit le tableau ci dessous avec les effectifs expérimentaux $n_{i,j}$, et les effectifs attendus $n_{i,j}^{th}$ si les deux variables A et B sont indépendantes. Pour cela, on calcule les totaux par lignes n_i' et les totaux par colonnes n_j''. Et on utilise la formule

$$n_{i,j}^{th} = \frac{n_i' \, n_j''}{n}.$$

A \ B	A_1	A_2	$\dots$	A_k	Totaux
B_1	$n_{1,1}$ / $n^{th}_{1,1}$	$n_{1,2}$ / $n^{th}_{1,2}$	$\dots$	$n_{1,k}$ / $n^{th}_{1,k}$	n'_1
B_2	$n_{2,1}$ / $n^{th}_{2,1}$	$n_{2,2}$ / $n^{th}_{2,2}$	$\dots$	$n_{2,k}$ / $n^{th}_{2,k}$	n'_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
B_l	$n_{l,1}$ / $n^{th}_{l,1}$	$n_{l,2}$ / $n^{th}_{l,2}$	$\dots$	$n_{l,k}$ / $n^{th}_{l,k}$	n'_l
Totaux	n''_1	n''_2	$\dots$	n''_k	n

- (H_0) : Les variables A et B sont indépendantes
- Conditions d'application : $n \geq 30$ et tous les effectifs théoriques sont supérieurs à 5
- $Z = \sum_{i=1}^{l}\sum_{j=1}^{k} \frac{(N_{i,j} - n^{th}_{i,j})^2}{n^{th}_{i,j}}$ suit une loi du χ^2 à $(l-1)(k-1)$ degré de liberté.

5.3 Test de Mann-Whitney-Willcoxon pour échantillons indépendants.

On cherche à comparer les valeurs x_i obtenues pour un échantillon à celles y_j obtenues pour un échantillon différent. On pourrait utiliser le test de comparaison de moyennes, mais lorsque les échantillons sont petits, ou les lois de X et Y ne peuvent pas vraiment être considérés comme normales,, on préfère utiliser le test de Mann-Whitney-Willcoson, basé sur les rangs. Voici la procédure à suivre :

- On inter-classe les données des deux échantillons **dans l'ordre croissant**. Cela donne le tableau ci-dessous.

Rang	1	2	3	$\dots$	$n_1 + n_2$
Valeur	x_3	y_1	y_4	$\dots$	x_5
Echantillon	1	2	2	$\dots$	1

- (H_0) : X et Y ont même la distribution (ou la même moyenne si leur loi est normale).
- On calcule la somme t_1 des rangs de l'échantillon 1.
- Si $n_1 + n_2$ est assez grand, on a

$$T_1 \sim \mathcal{N}(m_1, \sigma_1), \qquad m_1 = \frac{n_1(n_1+n_2+1)}{2}, \qquad \sigma_1 = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}$$

- Cela implique que $\frac{T_1 - m_1}{\sigma_1} \sim \mathcal{N}(0,1)$. La même chose marche pour T_2.