Notes cours Biostat L2

M. Bailly-Bechet

Université Claude Bernard Lyon 1 – France

Table des matières

Partie proba/stats du cours

DIAPOS 1-7 presentation module/notation/reussite

Des statistiques pour quoi faire ?

DIAPO 8 pourquoi stats

— Savoir si l'environnement a un effet sur le poids des pandas à la naissance
— Savoir si l'expression d'un gène peut faciliter le développement d'une tumeur
— Savoir si les acheteurs de céréales sont plus sensibles à la couleur de la boîte ou au prix d'achat

D'une manière générale, les statistiques permettent de répondre à ce type de question, *de manière quantifiée*, dans des situations mettant en jeu une certaine *variabilité*.

On peut artificiellement décomposer les statistiques en :

Statistique descriptive : la représentation graphique et le résumé de données observées à l'aide d'indice statistiques (*i.e.* la moyenne)

Statistique inférentielle : l'induction de propriétés d'une population à partir de données observées sur un échantillon.

DIAPO 9 lien stat desc et stats inferentielle

Plan du cours de stats : probas, puis généralités sur IC et tests, puis chi2, ANOVA, corrélation/régression et finalement comparaison de modèles.

1 Variables aléatoires et lois de probabilité

Une variable aléatoire est le résultat d'un tirage probabiliste. C'est une variable qui peut prendre plusieurs valeurs, avec des probabilités données.

En biologie, on observe des caractères sur les individus : ce sont des grandeurs qui peuvent prendre plusieurs états ou *modalités*

En statistiques, on travaille avec des variables aléatoires : ce sont des variables qui peuvent prendre plusieurs *valeurs* avec une certaine probabilité

Caractère biologique (couleur)	$\Leftrightarrow$	Variable aléatoire X
État (bleu, vert, rouge)	$\Leftrightarrow$	valeur x de probabilité $p(X = x)$

Les variables qualitatives sont les variables pour lesquelles une mesure est difficile à produire, ou subjective : couleur, type de régime alimentaire, intensité de la douleur...

Les variables quantitatives sont les variables que l'on peut mesurer explicitement : taille, poids, nombre de pattes...

Les variables quantitatives peuvent être distinguées par :
— leur espérance notée $\mathbb{E}(X)$ ou μ (valeur moyenne attendue). Une variable d'espérance 0 est dite *centrée*
— leur écart-type notée σ (variabilité attendue des résultats autour de la moyenne ; **exemple des notes des étudiants autour de 10**). Une variable d'écart-type 1 est dite *réduite*. On utilise souvent pour des raisons mathématiques σ^2 ou $\mathbb{V}(X)$, la variance.

On distingue :
— les variables quantitatives *discrètes*, ne pouvant prendre qu'un nombre fini de valeurs (par exemple le nombre de jambes d'un individu).
— les variables quantitatives *continues*, pouvant prendre un nombre infini de valeurs (par exemple la taille d'un individu).

1.1 Variables discrètes

La *loi de probabilité* d'une v.a. discrète est la probabilité de chaque résultat possible, notée $p(X = x)$. Si on lance 1 dés, la loi de probabilité D est :

s	$p(D=d)$
1	$\frac{1}{6}$
2	$\frac{1}{6}$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{1}{6}$
6	$\frac{1}{6}$

Peut-on prédire le résultat d'un dé ? Et pour deux dés, la somme ? Et le temps de demain ? Intution : plus il y a de variables, plus on peut prédire le résultat.

On a toujours, si les résultats possibles sont notés x_i avec $i = 1..N$, $\sum_{i=1}^{N} p(X = x_i) = 1$.

On a toujours

$$P(a \leq X \leq b) = \sum_{x=a}^{b} p(X = x).$$

Une loi discrète de probabilité : la loi binomiale La *Loi binomiale* est la loi d'une v.a. correspondant au nombre de succès lors du tirage de n variables de Bernouilli indépendantes. Chaque variable de Bernouilli est p succès $1 - p$ échec. On la note souvent $\mathcal{B}(n, p)$.

$$p(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{1}$$

$$\mathbb{E}(X) = np \tag{2}$$

$$\mathbb{V}(X) = np(1-p). \tag{3}$$

DIAPO 11 loi binomiale

Une loi discrète de probabilité : la loi de Poisson La *loi de Poisson* est la loi d'une v.a. correspondant au nombre d'evenements indépendants qui se produisent dans un intervalle donné, si leur fréquence est constante et connue (on la note λ). On la note souvent $\mathcal{P}(\lambda)$.

Exples : **mutations**, fréquence de passage d'un individu à un endroit

précis.

$$p(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{4}$$

$$\mathbb{E}(X) = \lambda \tag{5}$$

$$\mathbb{V}(X) = \lambda. \tag{6}$$

DIAPOS 12-14 loi théorique + représentation

1.2 Variables continues

$$f(x) = \frac{p(x)}{\Delta x},$$

avec Δx le pas que l'on voit.

DIAPOS 15-16 continu vers discret

La loi de probabilité d'une v.a. continue est donnée par sa *densité de probabilité*. Comme vu sur la diapo précédente pour une variable continue, $p(X = x) = 0$; on ne peut pas utiliser le formalisme du cas discret. La densité f associée à la variable aléatoire X est la probabilité de tirer une valeur dans un intervalle tout petit autour de x. On a toujours :

$$f(x) \geq 0 \quad \int_\Omega f(x) = 1.$$

On a toujours

$$P(a < X < b) = \int_a^b f(x)dx.$$

On note la similarité entre discret et continu en passant de $\sum$ à $\int$.

Un exemple de variable continue : la loi normale La loi normale est la loi de probabilité des variables aléatoires continues dépendantes d'un grand nombre de causes indépendantes et additives. Elle se note $\mathcal{N}(\mu, \sigma)$ avec μ l'espérance de la loi et σ l'écart-type. Attention à la notation de l'écart-type.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{7}$$

$$\mathbb{E}(X) = \mu \tag{8}$$

$$\mathbb{V}(X) = \sigma^2. \tag{9}$$

DIAPO 17 loi theorique

La loi de Student La loi de Student est une variante de la loi normale que l'on observe quand la variance de la variable étudiée est inconnue. On l'utilise toujours de manière indirecte ; elle dépend d'un nombre de *degrés de liberté* ; plus ce nombre est grand, plus elle est proche d'une loi normale de même moyenne et écart-type.

DIAPO 18 loi theorique
DIAPOS 19 representation variables

1.3 TCL et importance de la loi normale

Un énoncé du théorème central limite (TCL) est : *Toute somme de n variables aléatoires indépendantes converge vers une loi normale quand n devient grand.*

On déduit également *que la loi de la moyenne d'un échantillon est une loi normale.* Biologiquement, on en déduit que la somme de nombreuses causes indépendantes (par exemple de nombreux gènes – taille taille des mains chez l'homme –, de nombreux individus – quantité d'oxygène nette produite par une forêt –,. . .) est une loi normale. Pas mal de soucis dans la finance moderne viennent du fait qu'on fait des hypothèses avec des lois normales alors que les variables ne sont pas indépendantes ; exemple vente de Game of Thrones, les N tomes ne sont pas indépendants !

2 Rappels de statistiques descriptives ; estimation et intervalles de confiance

On rappelle qu'un échantillon est une sous-partie de la population étudiée.

L'objectif de l'*inférence* statistique consiste à trouver les valeurs de certaines caractéristiques de la population, à partir de celles observées dans l'échantillon.

Quand on veut la valeur numérique d'un paramètre, on parle d'*estimation.*

Une remarque importante est que l'inférence statistique ne dit pas si les choses sont ou ne sont pas dans la population, mais elle donne une probabilité à différents évenements, ou une probabilité à la valeur de certains paramètres.

2.1 Estimation ponctuelle

Dans ce cours, on peut vouloir estimer 3 paramètres dans une pop : la moyenne d'une variable μ, sa variance σ^2 et une fréquence théorique p.

On veut mesurer la durée de l'hibernation chez les marmottes, notée μ pour la population. On prend un échantillon de n marmottes pour lesquelless on chronomètre l'hibernation. On a une série statistique $x_1, ...x_n$. On peut calculer la moyenne de cette série. On rappelle que pour calculer la moyenne d'une série statistique, on a 2 formules :

Données non groupées :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{10}$$

Données groupées :

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} n_j x_j^\star, \text{ avec } n = \sum_{j=1}^{k} n_j \text{ et } x_j^\star \text{ la médiane de la classe } j. \tag{11}$$

DIAPOS 20-21 mangue avec moyenne

On dit que $\bar{x}$ est un estimateur de μ.

On peut montrer que $\bar{x} \to \mu$ quand $n \to \infty$: on dit que la moyenne empirique (observée) $\bar{x}$ est un estimateur non biaisé de μ. La meilleure estimation ponctuelle de μ que l'on puisse faire à partir des $x_1...x_n$ est $\hat{\mu} = \bar{x}$.

De la même manière, si on veut estimer la fréquence d'occurence d'un caractère comme un allèle particulier, on va compter, sur n marmottes, combien ont cet allèle. On note ce nombre k. Dans la population, la vraie probabilité d'avoir l'allèle en question est p; on peut montrer que $f = \frac{k}{n}$ est un estimateur non biaisé de p. On note $\hat{p} = f = \frac{k}{n}$

En ce qui concerne l'estimation de la variance σ^2 de la durée d'hibernation, un léger problème se pose. On rappelle qu'on peut calculer la variance observée ainsi :

Données non groupées :

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\,, \tag{12}$$

on développe et on obtient :

$$s^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2 \tag{13}$$

Sur des données groupées, par le même calcul, on a les deux formules :

$$s^2 = \frac{1}{n}\sum_{j=1}^{k} n_j\left(x_j^\star - \bar{x}\right)^2\,, \tag{14}$$

ou encore

$$s^2 = \frac{1}{n}\left(\sum_{j=1}^{k} n_j\left(x_j^\star\right)^2\right) - \bar{x}^2. \tag{15}$$

DIAPO 22 mangue avec moyenne

L'estimateur naturel serait s^2 ; mais cet estimateur est biaisé et sous-estime la variance globale dans la population (car on rate forcement les individus les plus extrêmes si on en prend peu). Il faut corriger cet estimateur ; un estimateur non biaisé de la variance de la population est $\hat{\sigma}^2 = \frac{n}{n-1}s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Bien faire la différence entre σ^2, variance de la pop, son estimateur $\hat{\sigma}^2$ et la variance observée s^2.

2.2 Distribution d'échantillonnage

A partir d'une population, on prend généralement un échantillon aléatoire. On pourrait en prendre plusieurs ; ils seraient différents, et les valeurs des variables mesurées dans chaque échantillon ne seront pas les même, et ne seront pas strictement identiques à celles de la population (sauf constance). On parle de *distribution d'échantillonnage* d'une variable. Grâce aux probabilités, on peut calculer cette distribution.

Que vaut cette distribution d'échantillonage ? Prenons le cas de la moyenne d'un grand échantillon. On a vu avec le TCL que la somme d'un grand nombre de v.a quelconques suit une loi normale. En particulier la moyenne observée $\bar{x}$ aura les caractéristiques suivantes :

Soit X une v.a. de moyenne μ et d'écart-type σ. Sa loi est inconnue ou qcq, on prend une loi uniforme comme exemple :

DIAPO 23 distro moyenne avec n variable mais tjs grand

— une espérance de μ
— une variance de $\frac{\sigma^2}{n}$
— suivra une loi normale, car elle est la somme d'une très grand nombre de variables indépendantes.

DIAPOS 24-28 repartition des valeurs autour de mu et sigma : 95%, 90%, 99.9%

Donc la notion de taille d'intervalle pour un risque donne de se planter. Exemples taille étudiants dans amphi d'à coté, notion d'erreur si je fais une prédiction trop précise ; a l'inverse notion que si je prends un risque ridicule je prédis une moyenne entre 1m et 3m !

Mathématiquement, on écrit, que, si on prend un risque α de se tromper :

$$P(\mu - C_\alpha < \bar{x} < \mu + C_\alpha) = 1 - \alpha$$
$$P(\mu - \epsilon_\alpha\sqrt{\frac{\sigma^2}{n}} < \bar{x} < \mu + \epsilon_\alpha\sqrt{\frac{\sigma^2}{n}}) = 1 - \alpha$$
$$P(\epsilon_\alpha < \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} < \epsilon_\alpha) = 1 - \alpha$$

Or $\frac{\bar{x}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$ est une v.a. centrée réduite qui suit une loi normale comme $\bar{x}$, car c'est un transformation linéaire d'une v.a. normale ; on peut donc trouver la valeur de ϵ pour un risque α indépendamment de μ et σ, en disant que :

$$P(\epsilon_\alpha < \mathcal{N}(0,1) < \epsilon_\alpha) = 1 - \alpha$$

DIAPO 29 loi normale centree reduite.

Ces valeurs seront les mêmes pour tous les problèmes où on se ramènera a une loi normale centree réduite, et donc tous les problèmes où on aura les mêmes hypothèses au départ. On les lit dans des tables (voir TD). Le seuil le plus couramment utilisé est $\epsilon_{0.05} = 1.96$.

DIAPO 30 table stat ecarts reduits.

On a donc au final :

$$P(\mu - \epsilon_\alpha\sqrt{\frac{\sigma^2}{n}} < \bar{x} < \mu + \epsilon_\alpha\sqrt{\frac{\sigma^2}{n}}) = 1 - \alpha.$$

2.3 Estimation par intervalle de confiance

L'idée de l'estimation par intervalle de confiance est d'associer à l'estimation ponctuelle la connaissance que l'on a sur la distribution d'échantillonnage. En fonction des situations et des hypothèses, on peut avoir une idée plus ou moins precise de la distribution d'échantillonnage, et donc un intervalle de confiance plus ou moins précis.

On construit l'IC au risque α de se tromper en regardant l'intervalle qu'on s'autorise à avoir dans la distro d'échantillonnage au risque α, et en appliquant cet intervalle autour de la valeur estimée. Dans le cas précédent, on va chercher à transformer l'expression que l'on a en un encadrement de μ, qui est inconnu et nous intéresse : FAIRE CALCUL EN FONCTION TEMPS

$$P(\mu - \epsilon_\alpha\sqrt{\frac{\sigma^2}{n}} < \bar{x} < \mu + \epsilon_\alpha\sqrt{\frac{\sigma^2}{n}}) = 1 - \alpha. \tag{16}$$

$$P(-\epsilon_\alpha\sqrt{\frac{\sigma^2}{n}} < \bar{x} - \mu < \epsilon_\alpha\sqrt{\frac{\sigma^2}{n}}) = 1 - \alpha. \tag{17}$$

$$P(-\bar{x} - \epsilon_\alpha\sqrt{\frac{\sigma^2}{n}} < -\mu < -\bar{x} + \epsilon_\alpha\sqrt{\frac{\sigma^2}{n}}) = 1 - \alpha. \tag{18}$$

$$P(\bar{x} + \epsilon_\alpha\sqrt{\frac{\sigma^2}{n}} > \mu > \bar{x} - \epsilon_\alpha\sqrt{\frac{\sigma^2}{n}}) = 1 - \alpha. \tag{19}$$

$$IC : \left[\bar{x} - \epsilon_\alpha\sqrt{\frac{\sigma^2}{n}}, \bar{x} + \epsilon_\alpha\sqrt{\frac{\sigma^2}{n}}\right] \tag{20}$$

On peut construire les IC pour d'autres hypothèses (voir cours sur Internet, bouquins biostatistiques à la BU, TDs). Le principal cas à connaître est quand la variance a été estimée à partir des données : Si X suit une *loi normale* de *variance inconnue* – **bien définir une variance inconnue** – on fait une petit erreur car on doit estimer la variance ; la loi sous-jacente n'est plus une loi normale mais une loi de Student, et l'IC devient :

$$IC : \left[\bar{x} - t_{\alpha,n-1}\sqrt{\frac{\hat{\sigma^2}}{n}}, \bar{x} + t_{\alpha,n-1}\sqrt{\frac{\hat{\sigma^2}}{n}}\right]$$

On peut trouver les t dans la table de Student, donnéee en TD ; leur valeur dépend à la fois de α et de n. On verra que si n est grand, $t_{\alpha,n-1} = \epsilon_\alpha$

Pour l'estimation d'une fréquence dans la population, la formule à connaitre, que l'on obtient par un raisonnement similaire, est :

$$IC : \left[\hat{p} - \epsilon_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + \epsilon_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right],$$

avec $\hat{p} = \frac{k}{n}$.

Si n est petit on ne peut pas faire grand chose. Les stateux veulent toujours un grand n.

toutes ces formules sont dans le formulaire distribué maintenant !

3 Tests

3.1 Raisonnement général des tests statistiques

On a les durées d'hibernation de n marmottes, notre échantillon. On peut faire un intervalle de confiance là dessus. Mais on peut également vouloir comparer ces valeurs à une moyenne de référence μ_0 (par exemple, le temps moyen d'hibernation des mêmes marmottes 10 ans plus tôt). Bien définir $\bar{x}$, μ et μ_0

Idée générique : si l'écart observé entre $\bar{x}$ et μ_0, moyenne de référence, est petit, on va dire que l'erreur est due au hasard ; si l'écart est grand on va dire que le hasard ne suffit pas. Cet écart va être calculé sous la forme de ce qu'on appelle la *statistique* du test.

Formellement, un test statistique distingue toujours 2 hypothèses :

H_0 l'hypothèse nulle : nos marmottes dorment autant que la moyenne de référence : la différence observée entre $\bar{x}$ et μ vient uniquement de la variabilité de la distribution d'échantillonage, donc du hasard. Mathématiquement on a $\mu = \mu_0$. Attention, $\bar{x} = \mu_0$ ne veut rien dire !

H_1 l'hypothèse alternative : le contraire, à savoir que nos marmottes ne dorment pas la même durée que la valeur de référence : il existe une différence réelle entre μ_0, la moyenne globale de la population, et μ, la moyenne de la sous-population de laquelle provient l'échantillon. Mathématiquement on a $\mu \neq \mu_0$. Attention, $\bar{x} \neq \mu_0$ ne veut rien dire !

Il faut remarquer que H_0 est structurellement plus simple que H_1, puisque H_0 implique qu'un seul paramètre décrit la population, alors que H_1 implique l'existance d'un deuxième paramètre. On dit que H_0 est l'hypothèse nulle

parce que c'est celle que l'on va privilégier (la plus simple) sauf si les données disent le contraire. Notion Rasoir d'Occam.

Logique des tests : les tests fonctionnent au rejet. Il faut se rappeler que $A \rightarrow B$ est équivalent à $nonB \rightarrow nonA$, mais pas du tout à $B \rightarrow A$. Exemple avec B : je mange toujours des céréales au petit déjeuner, et A : je suis un poulet. On a tjs $A \rightarrow B$. Si on ne mange pas de céréales ($nonB$), on peut en conclure que l'on n'est pas un poulet (si on en était un il faudrait qu'on mange des céréales). Mais on ne peut pas en conclure que manger toujours des céréales au petit dejeuner implique que vous etes un poulet, puisque d'autres choses que les poulets peuvent manger la meme chose qu'eux.

Pour chacun des échantillons ci-dessus, je peux réaliser un test statistique, qui va se baser sur l'assertion logique suivante : $H_0 \rightarrow$ statistique $\in$ [] (qui correspond à $A \rightarrow B$).

On va donc calculer la statistique.

1. Si elle est hors de l'intervalle, on a $nonB \rightarrow nonA$ et H_0 est fausse : on rejette H_0 si la statistique est forte.
2. Si la stat est dans l'intervalle, on ne peut pas en conclure logiquement que H_0 est vraie ; on va l'accepter par défaut, et parce que c'est l'hypothèse *la plus simple*.

La difficulté réside dans le fait qu' il n'y a pas de limite précise à l'intervalle qui nous intéresse, à cause des propriétés des lois statistiques qui nous intéressent : une loi normale peut donner n'importe quelle valeur, et même si H_0 est vraie, on peut observer – avec des probabilités différentes – n'importe quelle valeur de $\bar{x}$. La question est donc : quelle est la probabilité que $\bar{x}$ soit aussi éloigné de μ_0, si H_0 est vrai ?

DIAPOS 31-33 comparaison xbarre et seuils

On voit qu'en fonction du seuil de précision que l'on choisit, $\bar{x}$ est d'un cote ou de l'autre.

Comme précedemment, on va se ramener à une loi normale centré réduite. On a :

DIAPO 34 equivalence seuils

$$P(\bar{x}|\mathcal{N}(\mu, \sqrt{\frac{\sigma^2}{n}})) = P(\frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}|\mathcal{N}(0,1))$$

La question qui se pose est donc : si je prends comme hypothèse qu'un résultat au hasard doit tomber dans les $1 - \alpha$ pour cents des résultats les

plus probables, $\bar{x}$ est-il dans cet intervalle ?

La valeur critique correspondant à chaque risque α dépend des hypothèses du test effectué ; ici les valeurs critiques sont les mêmes ϵ_α que précedemment. En TD vous verrez les différents tests avec pour chacun, les hypothèses et les valeurs critiques correspondantes ; la démarche à se rappeler est toujours la suivante :

— Choisir un risque α ;
— En déduire en fonction du test la valeur seuil z_α.
— Calculer la statistique du test, z_{obs}.
— Si $\|z_{obs}\| \leq z_\alpha$, on est dans l'intervalle, on ne peut pas rejeter H_0.
— Si $\|z_{obs}\| > z_\alpha$, on est hors de l'intervalle, on peut rejeter H_0 et accepter H_1 ... avec un risque α de se tromper.

Une autre démarche s'est développée avec l'avènement de l'informatique : le calcul de la p-valeur. Cette valeur est la probabilité que H_0 explique bien les données observées : plus elle est faible, moins H_0 a de chances d'être vraie. Il faut cependant toujours choisir un niveau de risque avant de commencer les calculs, la seule différence vient de la méthode de calcul. Les p-valeurs ne se calculent pas à la main, mais toujours avec un ordinateur – voir R.

DIAPO 35 p valeur

Quand on effectue un test statistique avec un seuil choisi au risque α, on dit en pratique que si $\bar{x}$ appartient aux α pour cents de la distribution d'échantillonage les plus rares, on va rejeter H_0 comme étant fausse. Cette assertion est par définition fausse dans α pour cents des cas. On prend donc un risque α de se tromper, dit *risque de première espèce*.

Mais peut-on se tromper en choisissant H_0 aussi ?

Il existe un autre risque de se tromper : c'est celui ou on conserve H_0 par défaut alors que H_0 était fausse. C'est le cas ou H_1 est vraie, mais peut-etre pas tres différente de H_0, et donc on ne voit pas bien la différence. On note ce risque de *deuxième espèce* β. Dans la pratique ce risque est complexe à calculer, mais il est toujours présent.

DIAPO 36-37-38-39 exemple beta

Tableau recap risques

Réalité	H_0	H_1
Choix		
H_0	1-α	β
H_1	α	$1-\beta$

On ne peut pas minimiser à la fois α et β : si je minimise α, donc j'augmente

mes chances de conserver H_0 quand elle est vraie, alors je dois forcément augmenter β et augmenter ems chances de ne pas voir que H_1 est vraie...

Un mot sur la latéralité : si je m'intéresse à une hypothèse biologique unilatérale (un médicament par exemple), je vais changer mon seuil à l'avance, et décider que je ne considererai que les effets par exemple positifs. Dans ce cas, pour conserver le meme risque, il faut que je prenne un ϵ_α différent ; vu que la loi est symétrique, il faut que je prenne pour un test unilatéral un seuil $\epsilon_{2\alpha}$.

DIAPO 40 Test unilatéral

3.2 Différents types de tests

Il existe différents types de tests de comparaison de moyennes et de fréquences. En particulier, on peut vouloir comparer :

- Une moyenne observée à une moyenne de référence (test de conformité)
- Une fréquence observée à une fréquence ou probabilité de référence (conformité)
- L'égalité de 2 moyennes observées dans 2 échantillons différents (égalité ou homogénéité)
- L'égalité de 2 fréquences observées dans 2 échantillons différents
- L'égalité de 2 variances observées dans 2 échantillons différents

La procédure est toujours directe, sauf dans le cas où on veut comparer 2 moyennes observées. Dans ce cas, il faut d'abord vérifier si les variances des 2 populations desquelles viennent les 2 échantillons sont égales.

Si on a les durées d'hibernation d'un échantillon de marmottes des Alpes (n_A valeurs $x_1, x_2, ..., x_{n_A}$) et d'un échantillon de marmottes des Pyrénées (n_P valeurs $y_1, y_2, ..., y_{n_P}$), on doit :

- Faire un test pour vérifier l'égalité des variances ;
- Si ce premier test nous dit que les deux variances sont égales faire un test pour vérifier l'égalité des 2 moyennes.

Test de Fisher de comparaison de 2 variances On note s_A^2 la variance observée de la durée d'hibernation dans les Alpes, idem pour s_P^2. On note σ_A^2

et σ_P^2 respectivement les variances à l'échelle de la population.

$$H_0 : \sigma_A^2 = \sigma_P^2.$$
$$H_1 : \sigma_A^2 \neq \sigma_P^2.$$

On choisit un seuil $\alpha = 0.05$ par exemple. La valeur seuil de notre test sera alors lue dans la table de Fisher, et sera notée $F_{0.05}^{n_A-1,n_P-1}$. On appelle $n_A - 1$ et $n_P - 1$ les degrés de liberté. La statistique à calculer est :

$$F_{obs} = \frac{\hat{\sigma}_{max}}{\hat{\sigma}_{min}} = \frac{\hat{\sigma}_{P^2}}{\hat{\sigma}_{A^2}} = \frac{\frac{n_P s_P^2}{n_P-1}}{\frac{n_A s_A^2}{n_A-1}}, \tag{21}$$

si la variance observée est plus grande dans les Pyrénées comme ici. On voit que ce rapport devrait valoir une valeur proche de 1 si les variances observées sont proches, et donc que les variances des 2 populations sont supposément proches – ce qui est H_0.

On compare ensuite : si $F_{obs} \leq F_{0.05}^{n_A-1,n_P-1}$, on en conclut que H_0 ne peut pas être rejetée, et donc que les variances sont bien égales, avec un risque β de deuxième espèce inconnu. Si au contraire $F_{obs} > F_{0.05}^{n_A-1,n_P-1}$, on va rejeter H_0 avec un risque $\alpha = 5\%$ de se tromper, et dire que les variances sont différentes.

Test de comparaisons de 2 moyennes observées, variances égales Si les variances sont différentes, on ne peut pas tester l'égalité des moyennes ; si les variances sont égales, on peut faire le test, qui est alors direct. Brièvement :

Les variances étant considérées comme égales, on calcule la variance commune de nos deux échantillons :

$$\hat{\sigma}^2 = \frac{n_A s_A^2 + n_P s_P^2}{n_A + n_P - 2} \tag{22}$$

$$H_0 : \mu_A = \mu_P.$$
$$H_1 : \mu_A \neq \mu_P$$

On choisit un seuil $\alpha = 0.05$ par exemple ici aussi (c'est la valeur par défaut, et ca pose actuellement des problèmes). La valeur seuil de notre test

sera alors lue dans la table de Student, et sera notée $t_{0.05}^{n_A+n_P-2}$, encore une fois des degrés de liberté. La statistique à calculer est :

$$t_{obs} = \frac{|\bar{x} - \bar{y}|}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n_A} + \frac{1}{n_P}\right)}}, \tag{23}$$

On voit que ce rapport devrait valoir une valeur proche de 0 si les moyennes observées sont proches, et donc que les moyennes des 2 populations sont supposément proches, soit H_0.

On compare ensuite : si $t_{obs} \leq t_{0.05}^{n_A+n_P-2}$, on en conclut que H_0 ne peut pas être rejetée, et donc que les moyennes sont bien égales, avec un risque β de deuxième espèce inconnu. Si au contraire $t_{obs} > t_{0.05}^{n_A+n_P-2}$, on va rejeter H_0 avec un risque $\alpha = 5\%$ de se tromper, et dire que les moyennes sont différentes.

DIAPO 41-42 Comp moyennes marmottes + discussion p-valeur

Pour les formules détaillées de chaque test, formulaire, cours de première année, TDs et bouquins de biostats à la BU.

4 Test du χ^2

Le test du χ^2 est un test qui vise à analyser une table de contingence, c-à-d des comptes obtenus pour des variables qualitatives, discrètes ou regroupées par classe. Au sein de chaque groupe on a le *nombre* d'individus qui appartiennent au groupe.

DIAPOS 43-46 exemples chi2

4.1 χ^2 d'ajustement

On a une table de contingence pour une variable X. On se demande si les comptes observés pour chaque intervalle ou valeur de X suivent une loi donnée $p(X)$.

$$H_0 : X \text{ suit la loi } p(X)$$
$$H_1 : X \text{ ne suit pas la loi } p(X)$$

On note que suivre une loi connue est l'hypothèse nulle : c'est ce lle qui est structurellement plus simple, car une loi connue est plus précise que

"n'importe quelle autre loi". Attention au fait que cela peut paraître contre-intuitif ! On va voir si les données permettent de rejeter une hypothèse nulle disant qu'on suit bien une loi donnée.

Les étapes consistent en :

— Calculer les effectifs théoriques attendus si H_0 est vraie

— Regrouper les catégories pour que les effectifs théoriques vaillent au moins 5 (en réalité ne soient pas trop petits, on utilise une convergence vers la normale et pas la Poisson)

— Calculer la statistique qui mesure la différence entre effectifs théoriques et effetifs observés

— Conclure en comparant à la valeur théorique que l'on a pu tabuler si H_0 était vraie.

Si H_0 est vraie, l'effectif théorique de la classe i est donné par la formule $T_i = np(X = X_i)$, avec n l'effectif total.

X	X_1	X_2	...	X_k	Total
Effectifs observés O_i	n_1	n_2	...	n_k	n
Effectifs théoriques T_i	$np(X = X_1)$	$np(X = X_2)$	...	$np(X = X_k)$	n

La statistique est ensuite la suivante :

$$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(O_i - T_i)^2}{T_i}$$

On se rend compte dans la formule que l'on compare les effectifs théoriques et les effectifs observés ; si ceux-ci sont proches, la valeur du χ^2_{obs} sera faible, si les écarts sont grands la statistiques sera élevée.

On va comparer cette statistique à une valeur seuil χ^2_{k-1-c} à $k-1-c$ ddl, où c est le nombre de paramètres estimés. Le $n-1$ vient du fait que dans la somme de calcul du χ^2, on a n termes, mais le dernier est défini par les $n-1$ premiers, puisque on sait que la somme des effectifs totaux doit valoir n. De plus chaque paramètre estimé à partir des données donne artificiellement un meilleur ajustement, on enlève donc un ddl pour compenser ce biais : dans le cas ou on estime un seul paramètre, on peut en effet déduire le contenu des 2 dernières cases du tableau en sachant que la taille totale est n et que le paramètre estimé vaut la valeur calculée ; et ainsi de suite si on estime plus de un paramètre...

DIAPOS 47-48 ajustement à une loi de Poisson

4.2 χ^2 d'égalité

On a une table de contingence pour une variable X à k modalités mesurées dans m conditions. On se demande si les comptes observés pour chaque condition ont la même distribution.

DIAPO 49 Exemple Labos

H_0 : Les distributions sont les mêmes pour chaque condition

H_1 : Au moins une distribution est différente des autres pour une condition

La procédure va etre la même que plus haut, la différence venant de la table (plus complexe à première vue) et de la manière de calculer les effectifs théoriques.

A / B	A_1	A_2	$\dots$	A_p Somme	
B_1	n_{11}	n_{12}		n_{1p}	$n_{1\bullet}$
B_2	n_{21}	n_{22}		n_{2p}	$n_{2\bullet}$
$\dots$			$\dots$		
B_q	n_{q1}	n_{q2}		n_{qp}	$n_{q\bullet}$
Somme	$n_{\bullet 1}$	$n_{\bullet 2}$		$n_{\bullet p}$	n

Quel est l'effectif théorique dans la case ij ? Si H_0 est vraie, cet effectif est simplement proportionnel à l'effectif de la ligne i et de la conlonne j. On a donc :

$$t_{ij} = n\frac{n_{i\bullet}}{n}\frac{n_{\bullet j}}{n} = \frac{n_{i\bullet}n_{\bullet j}}{n}$$

Une fois ces effectifs théoriques calculés, on les regroupe pour avoir des cases supérieures a 5 si possible, puis on calcule le χ^2 comme précedemment :

$$\chi^2_{obs} = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{(n_{ij} - t_{ij})^2}{t_{ij}}$$

La valeur seuil du χ^2 dépend à la fois du risque de première espèce α et du nombre de degrés de liberté. Ce nombre vaut le nombre de cases indépendantes, sachant que les sommes sur les lignes et les colonnes sont

fixes ; on a donc $p-1$ colonnes indépendantes et $q-1$ lignes indépendantes, et $(p-1)(q-1)$ ddl et on a :

$$\chi^2_{seuil} = \chi^2_{\alpha,(p-1)(q-1)}$$

Si $\chi^2_{obs} \leq \chi^2_{seuil}$, on conserve H_0 par défaut, avec un risque β inconnu de se tromper ; si $\chi^2_{obs} > \chi^2_{seuil}$, on rejette H_0 et on accepte H_1 avec un risque α de se tromper.

4.3 χ^2 d'indépendance

En pratique, ce test ressemble énormément au précédent : on dispose de la table de contingence croisée pour un variable X affectée par 2 caractères A et B. Les hypothèses sont :

H_0 : Les caractères A et B sont indépendants

H_1 : Les caractères A et B ne sont pas indépendants

DIAPOS 50 Pb pandas

Les effectifs théoriques sont calculés de la même manière, le regroupement aussi, et le seuil de la même façon. La seule différence réside dans la formulation des hypothèses : les différentes modalités de A et B chacun sont-elles de simples variations ou des états complètement différents ? La différence n'est pas toujours évidente. Le nombre de ddl est le même que précedemment, pour les mêmes raisons.

DIAPO 51-52 Exemple Pandas + effectifs !

4.4 Lien entre test du χ^2 et test de comparison de proportions

A ZAPPER SI MANQUE DE TEMPS ET METTRE SUR SPIRAL

Dans les cas où on peut appliquer indifféremment un test du χ^2 ou un test de comparaison de proportions, les deux tests sont strictement équivalents. Par exemple prenons les données de réussite à un examen. On a un groupe d'étudiants avec leurs résultats, et on veut comparer à la moyenne nationale p.

(n'écrire au début que les effectifs observés).

X	Réussite	Échec	Total
Effectifs observés O_i	n_1	$n - n_1$	n
Effectifs théoriques T_i	np	$n(1-p)$	n

On a deux possibilités de test, pour les mêmes hypothèses H_0 et H_1 – l'hypothèse nulle testant le fait que les données sont réparties avec une proportion de réussite p.

— Le test de conformité d'une proportion observée à une proportion théorique. La statistique est :

$$\epsilon_{obs} = \left| \frac{\frac{n_1}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \right|$$

— Le test d'ajustement du χ^2 à une loi binomiale de paramètre p. **Ajouter partie tableau avec effectifs théoriques**. La statistique est :

$$\chi^2_{obs} = \frac{(n_1 - np)^2}{np} + \frac{((n - n_1) - n(1-p))^2}{n(1-p)}$$

Le lien entre ces deux formules est donné par le calcul suivant :

$$\begin{aligned}
\chi^2_{obs} &= \frac{(n_1 - np)^2}{np} + \frac{((n - n_1) - n(1-p))^2}{n(1-p)} \\
&= \frac{(n_1 - np)^2}{np} + \frac{(np - n_1)^2}{n(1-p)} \\
&= \frac{\left(\frac{n_1}{n} - \frac{np}{n}\right)^2}{\frac{n}{n^2}} \left(\frac{1}{p} + \frac{1}{1-p}\right) \\
&= \frac{\left(\frac{n_1}{n} - p\right)^2}{\frac{1}{n}} \left(\frac{1}{p(1-p)}\right) \\
&= \left(\frac{\frac{n_1}{n} - p}{\sqrt{\frac{p(1-p)}{n}}}\right)^2 = (\epsilon_{obs})^2
\end{aligned}$$

Si on regarde les valeurs seuils $\chi^2_{\alpha,1}$ et ϵ_α, on verra que l'on retrouve la relation $\chi^2_{\alpha,1} = \epsilon^2_\alpha$

5 ANOVA 1

On sait comparer les moyennes issues de 2 échantillons. Comment faire si l'on dispose de 3 échantillons $1, 2$ et 3 ? La première possibilité est de comparer :

— 1 et 2
— 2 et 3
— 1 et 3

Cela multiplie les tests, et peut conduire à des situations difficiles à interpréter : par exemple 1 et 2 ne sont pas significativement différents, 2 et 3 non plus, mais 1 et 3 le sont !.

L'objectif de l'ANOVA 1 est de tester simultanément l'égalité de toutes les moyennes de k échantillons. Chaque échantillon $i = 1..k$ est caractérisé par sa moyennes obervée $\bar{y}_i$ et sa variance observée s_i^2. Chaque échantillon est issu d'une population de moyenne μ_i et de variance σ_i^2. On veut donc tester :

$$H_0 : \mu_i = \mu_j \ \forall i, j$$
$$H_1 : \exists i, j \text{ t.q. } \mu_i \neq \mu_j.$$

DIAPOS 53-54-55 exemple donnees marmottes +graphique

Formellement, les données se présentent ainsi :

	A			
	A_1	A_2	...	A_p
y	y_{11}	y_{21}		y_{p1}
	y_{12}	y_{22}		y_{p2}
...			...	
	y_{1n_1}	y_{2n_2}		y_{pn_p}
Nombre de répétitions	n_1	n_2		n_p
Moyenne	$\bar{y}_1$	$\bar{y}_2$		$\bar{y}_p$
Écart-type observé	s_1^2	s_2^2		s_p^2

Le facteur A peut être qualitatif ou quantitatif, l'ANOVA peut toujours être effectuée – mais si le facteur est quantitatif on pourra faire mieux par la

suite. On dira que la taille totale de l'échantillon est $N = \sum_{i=1}^{p} n_i$. On note également la moyenne globale de tout l'échantillon $\bar{y} = \frac{1}{N} \sum_{i=1}^{p} \sum_{j=1}^{n_i} y_{ij}$.

Si on devait modéliser ces données, on pourrait le faire ainsi. Sous H_0 on peut écrire :

$$y_{ij} = \mu + e_{ij},$$

avec e_{ij} un terme de variabilité intrinsèque sur les mesures – on dira souvent que e_{ij}, qu'on appelle les résidus, suivent une loi normale de moyenne nulle. La moyenne théorique dans chaque groupe est donc bien μ, puisque le terme e_{ij} est aléatoire et n'ajoute rien à la moyenne.

Et sous H_1 on peut écrire :

$$y_{ij} = \mu + a_i + e_{ij}.$$

La différence est donc que sous H_1 on suppose que en plus du terme résiduel, on a un écart à la moyenne dans chaque groupe, avec $\mu_i = \mu + a_i$.

L'idee générale est que la variabilité des données autour de la moyenne globale, $\bar{y}$, est due à la fois à la variabilité au sein de chaque groupe, due au hasard, et à la variabilité moyenne entre les groupes, qui est nulle sous H_0 et vaut a_i pour le i$^{\text{ème}}$ groupe sous H_1. On va donc calculer ces deux variabilités et les comparer.

On va décomposer la variance globale :

$$\sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left[(y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}) \right]^2$$

$$\sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left[(y_{ij} - \bar{y}_i)^2 + (\bar{y}_i - \bar{y})^2 - 2 (y_{ij} - \bar{y}_i) (\bar{y}_i - \bar{y}) \right]$$

$$\sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{p} n_i (\bar{y}_i - \bar{y})^2 - 2 \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(y_{ij}\bar{y}_i - y_{ij}\bar{y} - \bar{y}_i^2 + \bar{y}_i\bar{y} \right)$$

$$\sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{p} n_i (\bar{y}_i - \bar{y})^2 - 2 \sum_{i=1}^{p} \left(n_i\bar{y}_i^2 - n_i\bar{y}_i\bar{y} - n_i\bar{y}_i^2 + n_i\bar{y}_i\bar{y} \right)$$

$$\sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{p} n_i (\bar{y}_i - \bar{y})^2$$

$$SCE_{tot} = SCE_{intra} + SCE_{inter}$$

Retour DIAPO 55 graphique marmotte pour SCE

On note :

$$\eta^2 = \frac{SCE_{inter}}{SCE_{tot}}$$

le rapport de la variabilité expliquée par des différences entre groupes (et donc par le facteur A) sur la variabilité totale. C'est un indicateur de la proportion de la variabilité qui est due au facteur A ; on a tjs $0 < \eta^2 < 1$.

On voit dans les formules que les différents SCE ne comprennent pas le même nombre de termes libres. Dans le SCE_{tot}, on utilise tous les y_{ij} ; ceux-ci sont tous indépendants sauf le dernier, on a donc $N-1$ ddl. Dans le terme SCE_{inter}, on utilise les $\bar{y}_i$: on a donc $p-1$ ddl. Dans le terme SCE_{intra}, on utilise les y_{ij} par rapport aux $\bar{y}_i$: on a donc $N-p$ ddl. On a :

$$N - 1 = p - 1 + N - p$$

On peut donc calculer à partir des SCE des carrés moyens, qui dépendent de ces ddl :

$$CM_{inter} = \frac{SCE_{inter}}{p-1} \quad CM_{intra} = \frac{SCE_{intra}}{N-p}$$

Si H_0 est vraie, on peut montrer que $CM_{intra} \sim CM_{inter}$, parce que la variabilité globale se décompose autant entre les groupes que dans les groupes, une fois la normalisation par les CM faite. Si H_0 est fausse alors on attend plus de variabilité entre groupes que dans chaque groupe (CM_{inter} plus fort que CM_{intra}). La statistique de l'ANOVA1 va donc être le rapport :

$$F_{obs} = \frac{CM_{inter}}{CM_{intra}}$$

Cette statistique, qui est en fait un rapport de variances, suit une loi de Fisher, et la valeur seuil est donc $F_{\alpha,p-1,N-p}$. Si $F_{obs} \leq F_{\alpha,p-1,N-p}$, on en conclut que H_0 ne peut pas être rejetée, et donc que les moyennes de tous les groupes sont bien égales, avec un risque β de deuxième espèce inconnu. Si au contraire $F_{obs} > F_{\alpha,p-1,N-p}$, on va rejeter H_0 avec un risque $\alpha = 5\%$ de se tromper, et dire que au moins une moyenne est différente des autres.

Notez bien qu'on parle d'analyse de variance pour comparer des moyennes – parce que la technique utilise une décomposition et un test basé sur les variances ; mais on compare bien des moyennes dans ce test.

DIAPOS 56-57 ANOVA marmotte

Cas particulier de 2 échantillons A ZAPPER SI MANQUE DE TEMPS ET METTRE SUR SPIRAL

La technique de l'ANOVA1 peut aussi s'appliquer à la comparaison de 2 échantillons. Prenons l'exemple simple de 2 échantillons de même taille n. La procédure classique de test d'égalité des moyennes, si les variances sont considérées comme égales, consisterait à calculer la statistique :

$$t_{obs} = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{1}{n}\right)}}.$$

Si on applique l'ANOVA 1 dans ces conditions, les calculs que l'on va effectuer sont les suivants :

$$\begin{aligned}
SCE_{inter} &= \sum_{i=1}^{p} n_i(\bar{y}_i - \bar{y})^2 = n\sum_{i=1}^{p}(\bar{y}_i - \bar{y})^2 = n\left[(\bar{y}_1 - \bar{y})^2 + (\bar{y}_1 - \bar{y})^2\right] \\
&= n\left[\left(\bar{y}_1 - \frac{\bar{y}_1 + \bar{y}_2}{2}\right)^2 + \left(\bar{y}_2 - \frac{\bar{y}_1 + \bar{y}_2}{2}\right)^2\right] = n\left[\frac{(\bar{y}_1 - \bar{y}_2)^2}{2}\right] \\
SCE_{intra} &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 = \sum_{j=1}^{n}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n}(y_{2j} - \bar{y}_2)^2 \\
&= ns_1^2 + ns_2^2 = \hat{\sigma}^2(2n-2) \\
F_{obs} &= \frac{CM_{inter}}{CM_{intra}} = \frac{\frac{SCE_{inter}}{1}}{\frac{SCE_{intra}}{2n-2}} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{\hat{\sigma}^2\left(\frac{2}{n}\right)} = t_{obs}^2
\end{aligned}$$

Quelques éléments pratiques : Les formules des SCE données plus haut sont justes, mais comme dans le cas des calculs de variance, il existe des formules développées plus simples. Ces formules développées s'obtiennent de la même manière que les formules développées dans le cas du calcul de la variance pour un échantillon. Les formules sont :

$$SCE_{tot} = \left(\sum_{i=1}^{p}\sum_{j=1}^{n_i} y_{ij}^2\right) - \frac{T^2}{N} \text{ avec } T = \sum_{i=1}^{p}\sum_{j=1}^{n_i} y_{ij} = N\bar{y}$$

$$SCE_{inter} = \left(\sum_{i=1}^{p} \frac{T_i^2}{n_i}\right) - \frac{T^2}{N} \text{ avec } T_i = \sum_{j=1}^{n_i} y_{ij} = n_i\bar{y}_i$$

$$SCE_{intra} = SCE_{tot} - SCE_{inter} = \left(\sum_{i=1}^{p}\sum_{j=1}^{n_i} y_{ij}^2\right) - \left(\sum_{i=1}^{p} \frac{T_i^2}{n_i}\right)$$

Les conditions d'application de ce test sont :

— Indépendance des différent échantillons (pas d'individus dans 2 échantillons) – supposée

— Normalité de la distribution de chaque échantillon (on parle parfois de normalité des résidus) – supposée mais testable avec le χ^2.

— Homoscédasticité des échantillons, ie égalité des variances.

Pour tester cette dernière condition, on ne peut pas employer les test de Fisher classique car on a plus de 2 variances. On a plusieurs cas :

— Si les n_i sont différents entre eux, le test exact est le test de Bartlett, non étudié cette année. On supposera alors l'homoscédasticité des données.

— Si les n_i sont tous identiques, on peut faire un test dit de Hartley.

La procédure du test de Hartely est la suivante :

$$H_0 : \sigma_i^2 = \sigma_j^2 \, \forall i, j$$
$$H_1 : \exists i, j \text{ t.q. } \sigma_i^2 \neq \sigma_j^2.$$

On calcule la statistique :

$$H_{obs} = \frac{s_{max}^2}{s_{min}^2},$$

Et on compare à la valeur seuil au risque α dans la table de Hartley. Ce tableau a deux entrées : la taille des groupes n_i, et le nombre de groupes comparés.

Pour éviter les calculs lourdingues, R

DIAPOS 58-59 ANOVA marmottes R

6 ANOVA2

Exemple : on veut étudier des données concernant la vitesse de réplication d'un virus de la grippe en fonction de la souche et de la température.

DIAPOS 60-61-62 présentation données virus + graphique

Formellement les données se présentent de cette façon :

	Facteur A			
Facteur B	A_1	A_2	...	A_p
B_1	y_{111}	y_{211}	$\cdots$	y_{p11}
	y_{112}	y_{212}	$\cdots$	y_{p12}
	$\cdots$	$\cdots$	$\cdots$	
	$y_{11n_{11}}$	$y_{21n_{21}}$	$\cdots$	$y_{p1n_{p1}}$
B_2	y_{121}	y_{221}	$\cdots$	y_{p21}
	y_{122}	y_{222}	$\cdots$	y_{p22}
	$\cdots$	$\cdots$	$\cdots$	
	$y_{12n_{12}}$	$y_{22n_{22}}$	$\cdots$	$y_{p2n_{p2}}$
...				
B_q	y_{1q1}	y_{2q1}	$\cdots$	y_{pq1}
	y_{1q2}	y_{2q2}	$\cdots$	y_{pq2}
	$\cdots$	$\cdots$	$\cdots$	
	$y_{1qn_{1q}}$	$y_{2qn_{2q}}$	$\cdots$	$y_{pqn_{pq}}$

L'ANOVA comme le choix du meilleur modèle Si toutes les données avaient la même moyenne, on aurait comme modèle sous-jacent :

$$y_{ijk} = \mu + \epsilon_{ijk},$$

avec ϵ_{ijk} un terme de bruit gaussien – variabilité suivant une loi normale centrée, dont l'écart-type est la variabilité typique des données. Si le facteur A a un effet particulier, le modèle devient :

$$y_{ijk} = \mu + a_i + \epsilon_{ijk}.$$

De même si le facteur B a un effet particulier, le modèle devient :

$$y_{ijk} = \mu + a_i + b_j + \epsilon_{ijk}.$$

Finalement, si la valeur de a_i dépend de la valeur de b_j, ou inversement, le modèle complet sous-jacent est le suivant :

$$y_{ijk} = \mu + a_i + b_j + c_{ij} + \epsilon_{ijk}.$$

La question que pose l'ANOVA2 – qui est une généralisation de la question posée par l'ANOVA1 – est de savoir quel est le meilleur modèle pour décrire les données. Les modèles avec plus de coefficients sont mathématiquement plus compliqués : ils seront des hypothèses alternatives dans les tests, les modèles les plus simples étant à chaque fois des hypothèse nulles.

L'ANOVA2 teste 3 hypothèses en parallèle :

— Sur le facteur A :

H_0 : Les moyennes dans les différentes categories du facteur A sont les memes.

H_1 : Les moyennes dans les différentes categories du facteur A sont differentes.

ou encore

$H_0 : a_i = 0 \, \forall i$

$H_1 : \exists i$ t.q. $a_i \neq 0$

— Sur le facteur B :

H_0 : Les moyennes dans les différentes categories du facteur B sont les memes.

H_1 : Les moyennes dans les différentes categories du facteur B sont differentes.

ou encore

$H_0 : b_j = 0 \, \forall j$

$H_1 : \exists j$ t.q. $b_j \neq 0$

— Sur l'interaction entre ces 2 facteurs :

H_0 : Les moyennes dans les différentes categories du facteur A dependent des valeurs de B.

H_1 : Les moyennes dans les différentes categories du facteur A ne dependent pas des valeurs de B.

ou encore

$H_0 : c_{ij} = 0 \, \forall i, j$

$H_1 : \exists i, j$ t.q. $c_{ij} \neq 0$

De la même manière que dans l'ANOVA1, on va comparer les variabilités dues aux différents facteurs entre elles. On va décomposer la variabilité globale en une somme, normaliser chaque terme par le nombre de *ddl* approprié, et comparer ces termes entre eux.

On suppose le nombre de répétitions par case n_{ij} égal dans toutes les cases, et on le note n. Le cas ou le nombre de répétitions est différent est en pratique calculatoirement complexe, et empêche de faire la décomposition ci-dessous, ce qui est problématique à la fois en terme d'interprétation et en

termes de calculs.

$$SCE_{tot} = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(y_{ijk} - \bar{y})^2$$

La décomposition employée est la suivante :

$$y_{ijk} - \bar{y} = (\bar{y}_{i\bullet} - \bar{y}) + (\bar{y}_{\bullet j} - \bar{y}) + (\bar{y}_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}) + (y_{ijk} - \bar{y}_{ij}).$$

$$\begin{aligned}
SCE_{tot} &= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(y_{ijk} - \bar{y})^2 \\
&= qn\sum_{i=1}^{p}(\bar{y}_{i\bullet} - \bar{y})^2 + pn\sum_{j=1}^{q}(\bar{y}_{\bullet j} - \bar{y})^2 + n\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y}_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y})^2 \\
&+ \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij})^2 \\
&= SCE_A + SCE_B + SCE_{A\times B} + SCE_{res}
\end{aligned}$$

Les *ddl* correspondants sont :

— Total : $N - 1$ avec $N = npq$.
— $A : p - 1$.
— $B : q - 1$.
— $A \times B : (p-1)(q-1)$ car on fait une somme de pq termes avec les moyennes de chaque catégorie fixées pour les p modalités de A et les q modalités de B.
— $Res : npq - pq = pq(n-1)$

Les calculs à faire sont donc :

$$CM_A = \frac{SCE_A}{p-1}$$
$$CM_B = \frac{SCE_B}{q-1}$$
$$CM_{A\times B} = \frac{SCE_{A\times B}}{(p-1)(q-1)}$$
$$CM_{res} = \frac{SCE_{res}}{pq(n-1)}$$

On va ensuite répondre aux 3 tests en calculant les valeurs suivantes :

$$F_A = \frac{CM_A}{CM_{res}} \quad F_{seuil} = F^{\alpha}_{p-1,pq(n-1)}$$
$$F_A = \frac{CM_B}{CM_{res}} \quad F_{seuil} = F^{\alpha}_{q-1,pq(n-1)}$$
$$F_A = \frac{CM_{A\times B}}{CM_{res}} \quad F_{seuil} = F^{\alpha}_{(p-1)(q-1),pq(n-1)}$$

On conclut de la manière habituelle.

DIAPOS 63-65 exemple bio
DIAPOS 66-70 deuxième cas avec interaction

Détails pratiques Si le plan est déséquilibré, les calculs précédents sont invalides. On fait sur machine, et attention, le problème n'est plus symétrique : on "attribue" la variance préférentiellement à A ou B, voir TP R sur ANOVA2.

Comme pour l'ANOVA1, on emploie en TD les formules développées (équivalence en 2 lignes avec les formules précedentes) :

$$SCE_{tot} = \left(\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n} y_{ijk}^2\right) - \frac{T^2}{N} \qquad T = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n} y_{ijk}$$
$$SCE_A = \frac{1}{qn}\left(\sum_{i=1}^{p} T_{i\bullet}^2\right) - \frac{T^2}{N} \qquad T_{i\bullet} = \sum_{j=1}^{q}\sum_{k=1}^{n} y_{ijk}$$
$$SCE_B = \frac{1}{pn}\left(\sum_{j=1}^{q} T_{\bullet j}^2\right) - \frac{T^2}{N} \qquad T_{\bullet j} = \sum_{i=1}^{p}\sum_{k=1}^{n} y_{ijk}$$
$$SCE_{res} = \left(\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n} y_{ijk}^2\right) - \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} T_{ij}^2 \qquad T_{ij} = \sum_{k=1}^{n} y_{ijk}$$
$$SCE_{A\times B} = SCE_{tot} - SCE_A - SCE_B - SCE_{res}$$

Les conditions d'application sont :

— Indépendance des différent échantillons (pas d'individus dans 2 échantillons) – supposée

— Normalité de la distribution de chaque échantillon (on parle parfois de normalité des résidus) – supposée mais testable avec le χ^2.

— Homoscédasticité des échantillons, ie égalité des variances – on vérifie avec Hartley si les n_{ij} sont égaux.

Cas particulier : $n = 1$ Si on n'a pas de répétitions, on peut voir que les formules précédentes ne permettent pas de calculer un SCE_{res} : on a $SCE_{res} = 0$. Donc on ne peut pas savoir s'il y a ou non interaction et calculer en même temps les résidus. SCE_{res} étant la valeur de référence pour tous les tests, cette absence nous oblige à considérer qu'on ne peut pas dans ce cas distinguer l'effet d'interaction et l'effet résiduel, et on va décomposer la variance en un terme du à A, un terme du à B, et un unique terme résiduel. On va donc procéder comme pour une ANOVA2 classique avec les modifications suivantes :

$$SCE_{tot} = SCE_A + SCE_B + SCE_{res}$$
$$SCE_{res} = SCE_{tot} - SCE_A - SCE_B$$

Les 2 premiers tests restent inchangés ; le test d'interaction n'est plus réalisable.On est obligé d'ajouter dans les hypothèse de départ qu'il n'y a pas interaction – s'il y en a une les résultats seront faussés, puisqu'on modélise implicitement que les $c_{ij} = 0$ dans ce test.

7 Analyse bivariée

DIAPOS 71-72 ribosome et présentation problème

Beaucoup d'expériences, en biologie notamment, mènent à considérer simultanément deux variables X et Y *appariées*, c'est-à dire où à chaque individu de l'échantillon correspond une valeur de X et une valeur de Y. On peut :

Décrire et quantifier les relations entre deux variables : est ce que la concentration en ARNm (X) dans la cellule et la concentration dans la protéine correspondante (Y) sont liées et est-ce que cette liaison est linéaire ? C'est un calcul de *corrélation*.

Modéliser pour prédire les valeurs de Y à partir des valeurs de X : connaissant X, que puis-je dire pour Y ? C'est un calcul de *régression*.

Si elle existe, la variable *contrôlée* X est appelée variable *indépendante* ou *explicative*, et est toujours en abcisse. La variable aléatoire Y est appelée variable *dépendante* ou *à expliquer*, et est toujours placée en ordonnée. Si on a 2 variables non contrôlées, le sens du graphe n'est pas prédéterminé, mais il est souvent implicite que X est la cause de Y.

DIAPOS 73-74 graphique ribosome+marmotte controle

Si l'on a plus de 2 variables que l'on veut analyser simultanément, on procèdera à une analyse *multivariée*.

7.1 Covariance et coefficient de corrélation linéaire

On prend 2 variables X et Y appariées. On note x_i et y_i, $i = 1..n$ les valeurs prises dans les échantillons. On note $\bar{x}$ et s_X^2 la moyenne et la variance observées de X, idem pour Y.

On définit la *covariance* de deux variables aléatoires X et Y :

$$\mathrm{cov}(X, Y) = \sigma_{XY} = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y),$$

avec $\mathbb{E}$ le symbole de l'espérance d'une variable aléatoire. Notez le lien avec les formules de la variance – la variance d'une v.a. est la covariance d'une variable avec elle-même :

$$\text{var}(X) = \text{cov}(X, X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2, \tag{24}$$

$$\text{à lier à } s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2. \tag{25}$$

À partir de deux échantillons de tailles n, on peut mesurer la *covariance observée* :

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \bar{x}\bar{y}$$

On a alors la covariance estimée de la population :

$$\hat{\sigma}_{XY} = \frac{n}{n-1} s_{XY}$$

À partir de la covariance, et des écarts-types de X et Y, on peut définir le *coefficient de corrélation linéaire* ou coefficient de Pearson de deux variables aléatoires :

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y},$$

avec σ_X et σ_Y les écarts-types de X et Y respectivement. Cette valeur mesure à quel point X et Y varient ensemble – on voit que la variable sera grande si les x_i et les y_i sont simultanément au dessus de leurs moyennes respectives, ou en dessous.

DIAPO 75 graphe en 4 parties

Cette valeur est la vraie valeur du coefficient de corrélation linéaire, une valeur à laquelle on ne peut avoir accès que partiellement au travers des échantillons qu'on a :

$$\hat{\rho} = r_{XY} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y} = \frac{\frac{n}{n-1} s_{XY}}{\sqrt{\frac{n}{n-1}} s_X \sqrt{\frac{n}{n-1}} s_Y} = \frac{s_{XY}}{s_X s_Y},$$

On emploie plus souvent la notation r_{XY} que $\hat{\rho}$ dans la pratique, pour des raisons historiques. Ce coefficient mesure à quel point les variables X et Y suivent une relation linéaire.

DIAPO 76 graphe coeff correlation

r_{XY} est toujours compris entre les 2 extrêmes d'alignement "parfait" ; on a :

$$-1 \leq r_{XY} \leq 1.$$

7.2 Test du coefficient de corrélation

Le mesure du coeff de corrélation linéaire ne dit pas si la liaison observée est due au hasard de l'échantillonage ou à une réelle liaison entre variables. Par exemple deux points au hasard vont toujours être alignés, et si on a peu de points il va être difficile de juger si l'alignement est aléatoire ou pas. Pour tester cette hypothèse, on va effectuer un test statistique.

Pour pouvoir être appliqué, il faut que les deux variables X et Y soient distribuées normalement. Si ce n'est pas le cas, il faudra employer un test non paramétrique, le test de corrélation de Spearman (cours MAB en L3). Plus généralement, si les variables ne semblent pas distribuées normalement (nuage de points elliptique), l'usage du coefficient de corrélation pour mesurer la liaison entre variables est dangereux et peut conduire à de fausses conclusions.

DIAPO 77 graphe Anscombe

Le test du coefficient de corrélation linéaire entre deux variables étudiées X et Y, a pour hypothèses :
$H_0 : \rho = 0$, X et Y sont linéairement indépendantes
$H_1 : \rho \neq 0$, X et Y sont linéairement dépendantes

Le test est un test de Student. La statistique à calculer, pour des échantillons de taille n, est :

$$t_{obs} = \left| \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \right|$$

La valeur t_{obs} est ensuite comparée à la valeur seuil lue sur la table de Student, pour un risque α choisi à l'avance et $n-2$ degrés de liberté. Si $t_{obs} < t_{seuil}$, on ne peut pas rejetter l'hypothèse nulle, dans le cas contraire, on pourra accepter l'hypothèse alternative.

On voit que plus r est grand en valeur absolue, et plus le nombre de points augmente, plus on peut conclure à la significativité de la relation. On voit aussi que l'on ne peut pas tester la significativité de la relation

entre 2 points seulement. À l'inverse, si on fait un test avec énormément de points (en génomique, classiquement 10 à 20000), il arrive qu'un coefficient de corrélation $r = 0.02$ soit significativement différent de 0. Qu'en penser ? On verra la signification de r comme pourcentage de la variation expliquée dans le chapitre sur la régression. Il ne faut pas confondre *taille d'effet* et *taille d'échantillon*, l'acceptation de H_1 dans le test venant toujours d'un mélange des deux.

7.3 Exemples

DIAPOS 78-79-80 Exple concentration ARNm + erreur R cov
DIAPO 81 Exple Pandas

Dans le cas où on a des données numériques groupées, on emploie les formules de la moyenne, l'écart-type et la covariance pour des données groupées par classe :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^{'},$$

$$s_X^2 = \frac{1}{n}\left(\sum_{i=1}^{p} n_i x_i^{'}\right) - \bar{x}^2,$$

$$s_{XY} = \frac{1}{n}\left(\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} x_i^{'} y_i^{'}\right) - \bar{x}\bar{y},$$

où les valeurs $x_i^{'}$ et $y_i^{'}$ sont les médianes des valeurs de chaque classe.

On note bien la différence entre le test du χ^2 vu précedemment et le test de régression linéaire : les hypothèses testées ne sont pas les mêmes, le test de corrrélation teste une hypothèse bcp plus précise.

DIAPO 82-83 Calcul Pandas

8 Régression et modèle linéaire

8.1 Le modèle linéaire

On cherche à étudier l'évolution d'une variable Y en fonction d'une variable X, aléatoire ou contrôlée. La droite de régression linéaire est un modèle

de la relation entre une variable X et une variable Y par une droite, qui permet de *prédire* les valeurs de Y en fonction des valeurs de X.

La principale différence conceptuelle avec le cas où les 2 variables sont aléatoires est que, ici, on ne cherche pas à savoir s'il existe ou non une relation, mais plutôt quelle est la nature de cette relation. Si le test de corrélation a indiqué qu'il n'y avait pas de relation, alors écrire un modèle ne sert à rien...

Attention, l'existence d'une corrélation n'implique pas forcément celle d'une causalité. On a par exemple une très bonne corrélation entre le nombre de cigognes en Alsace et le taux de fertitilité en Asace, et pourtant les 2 phénomènes ne sont reliés causalement qu'au travers d'une cause commune : le passage du temps...

Si l'on veut modéliser une relation non-linéaire entre Y et X, on parlera de régression polynomiale, ou exponentielle, ou logarithmique en fonction de la fonction utilisée : c'est de la régression *non linéaire*.

DIAPOS 84-85-86 choix droite regression et ecarts

Valeurs observées : (x_i, y_i)
Valeurs prédites : $\hat{y}_i = ax_i + b$
Ecarts : $e_i = y_i - \hat{y}_i$

8.2 Estimation des paramètres

On estime les paramètres a et b en trouvant les valeurs qui minimisent les écarts entre $\hat{y}_i$ et y_i.

On note s_R^2 la variabilité résiduelle , t.q $s_R^2 = \sum_{i=1}^{n} e_i^2$. Cette variabilité est observée, on peut calculer une estimation de cette variabilité si on avait toute la population, et donc qu'on ne faisait aucune erreur sur a et b : $\hat{\sigma}_R^2 = \frac{n}{n-2} s_R^2$. Cette estimation ne nous sert pas pour le calcul des paramètres, car minimiser s_R^2 et σ_R^2 est équivalent.

On veut minimiser :

$$s_R^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (ax_i + b - y_i)^2,$$

en fonction des valeurs de a et b. Pour cela, on cherche quand les dérivées de

cette fonction par rapport à b et a sont nulles.

$$\frac{\partial F}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} (2b + 2ax_i - 2y_i) = 0, \tag{26}$$

$$2nb + 2an\bar{x} - 2n\bar{y} = 0, \tag{27}$$

$$b = \bar{y} - a\bar{x}. \tag{28}$$

$$\frac{\partial F}{\partial a} = 0 \Rightarrow \sum_{i=1}^{n} \left(2ax_i^2 + 2bx_i - 2x_i y_i\right) = 0, \tag{29}$$

$$\sum_{i=1}^{n} \left(2ax_i^2 + 2\bar{y}x_i - 2a\bar{x}x_i - 2x_i y_i\right) = 0, \tag{30}$$

$$a \sum_{i=1}^{n} \left(x_i^2 - \bar{x}x_i\right) = \sum_{i=1}^{n} \left(x_i y_i - \bar{y}x_i\right), \tag{31}$$

$$a \left(\left(\sum_{i=1}^{n} x_i^2 \right) - n\bar{x}^2 \right) = \sum_{i=1}^{n} (x_i y_i) - n\bar{y}\bar{x}, \tag{32}$$

$$a = \frac{s_{XY}}{s_X^2}. \tag{33}$$

DIAPO 87 exemple calcul + 88 R

Notre *modèle* nous donne :

— des valeurs prédites, qui peuvent être comparées à la réalité.

— des valeurs pour les paramètres a et b, qui peuvent être interpretées et réemployées par la suite.

Ici, $\hat{a}$ est le taux d'augmentation de concentration en protéine par unité d'ARNm, et $\hat{b}$ représente la concentration de base en protéine si $[ARNm] = 0$ dans notre modèle. On note que $b \neq 0$, ce qui est biologiquement irréaliste (ou alors on doit tenir compte de ce qui a été passé à la cellule à sa naissance). On aurait pu forcer $b = 0$ en changeant le calcul précédent ; on a alors un modèle différent, et donc des paramètres différents. Attention aux unités dans notre modèle !

DIAPOS 89-90-91-92-93 residus

Intervalles de confiance Dans un cadre prédictif, plutôt que d'effectuer des tests, on peut vouloir écrire des intervalles de confiance au risque α autour

de la pente et de l'ordonnée à l'origine prédite. Les intervalles de confiance ont les formes suivantes (on ne rentre pas dans le pourquoi) :

$$IC_\alpha \quad a : \hat{a} \pm t_{\alpha,n-2}\sqrt{\frac{\hat{\sigma}_r^2}{ns_X^2}}$$

$$IC_\alpha \quad b : \hat{b} \pm t_{\alpha,n-2}\sqrt{\frac{\hat{\sigma}_r^2}{n}},$$

avec $\hat{\sigma}_R^2 = \frac{n}{n-2}\sum_{i=1}^n e_i^2$, comme vu avant. Un des intervalle de confiance les plus utiles est celqui que l'on peut former autour d'une prédiction : connaissant la régression linéaire de Y en fonction de X, quelle valeur y_0 peut-on espérer obtenir pour une valeur x_0 de la variable X ?

$$IC_\alpha \; y_0 : \hat{a}x_0 + \hat{b} \pm t_{\alpha,n-2}\sqrt{\hat{\sigma}_r^2\left(\frac{n+1}{n} + \frac{(x_0-\bar{x})^2}{ns_X^2}\right)}.$$

DIAPOS 94-95-96 IC et generalisation

On remarque que la précision de prédictions pour des x_0 situés loin de la valeur moyenne $\bar{x}$ sont beaucoup moins précises que celles correspondants à des valeurs proches. Attention aux extrapolations ! Les causes de variation de Y effectives dans l'intervalle étudié de X ne sont pas forcément vraies en dehors de cet intervalle. Il y donc deux sources d'erreurs quand on s'éloigne des valeurs de l'expérience : la fait que la variabilité statistique augmente, et le fait que le modèle peut etre biologiquement faux loin des valeurs expérimentales.

Tests Il existe également un test permettant de vérifier l'égalité d'une pente observée à une pente théorique, et ainsi de vérifier si des données expérimentales infirment ou non un modèle précédent. Le test est fait ainsi :

H_0 : la pente réelle a et la pente théorique γ sont égales.
H_1 : la pente réelle a et la pente théorique γ sont différentes.

La statistique à calculer est :

$$t_{obs} = \frac{|\hat{a} - \gamma|}{\sqrt{\frac{\hat{\sigma}_r^2}{ns_X^2}}},$$

Ce t_{obs} est à comparer à un t_{seuil} à $n-2$ ddl.

Dans le cas $\gamma = 0$, on montrera au prochain cours que ce test est équivalent au test du coeff de corrélation.

DIAPO 97 test conformite pente

Pour vérifier l'égalité de deux pentes réelles à partir de deux pentes mesurées, le test se déroule ainsi :

H_0 : les deux pentes réelles a_1 et a_2 sont égales.
H_1 : les deux pentes réelles a_1 et a_2 sont différentes.

La statistique à calculer est :

$$t_{obs} = \frac{|\hat{a}_1 - \hat{a}_2|}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n_1 s^2_{X,1}} + \frac{1}{n_2 s^2_{X,2}} \right)}},$$

avec :

$$\hat{\sigma}^2 = \frac{(n_1 - 2)\hat{\sigma}^2_{r,1} + (n_2 - 2)\hat{\sigma}^2_{r,2}}{n_1 + n_2 - 4}.$$

On compare ce t_{obs} à un t_{seuil} à $n_1 + n_2 - 4$ ddl.

9 Comparaisons de modèles

DIAPOS 98-99 présentation problème marmottes.

On peut, comme dans l'ANOVA1, décomposer la variance dans la régression linéaire, en se servant des valeurs prédites au lieu des moyennes des groupes comme point intermédiaire :

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \tag{34}$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{35}$$

$$SCE_{tot} = SCE_{res} + SCE_{y,x} \tag{36}$$

$$n s^2_Y = n s^2_r + n s^2_{y,x} \tag{37}$$

La somme des carrés totale se décompose en une somme *expliquée* par la régression et une somme *résiduelle*. Pour un modèle linéaire, la variabilité

se décompose en une part expliquée par la relation entre Y et X, et une part résiduelle qui est indépendante de X. Cette somme résiduelle est, si l'on regarde sa définition, la somme des carrés des résidus : il s'agit donc, à un facteur près, du s_R^2 vu au chapitre précédent !

De la même manière que pour une ANOVA 1 on définissait un rapport de corrélation pour donner la part de variabilité totale expliquée par le facteur :

$$\eta^2 = \frac{SCE_{inter}}{SCE_{tot}},$$

on définit pour le modèle de régression linéaire la part de variabilité linéairement expliquée par la variable X :

$$R^2 = \frac{SCE_{y,x}}{SCE_{tot}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n \left(ax_i + b - (a\bar{x} + b)\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \tag{38}$$

$$= \frac{\sum_{i=1}^n a^2 (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \tag{39}$$

$$R^2 = a^2 \frac{s_X^2}{s_Y^2}. \tag{40}$$

Or, on a $a = \frac{s_{XY}}{s_X^2}$; en remplacant on obtient :

$$R^2 = \left(\frac{s_{XY}}{s_X^2}\right)^2 \frac{s_X^2}{s_Y^2} = \frac{s_{XY}^2}{s_X^2 s_Y^2} = (r_{XY})^2$$

On a donc au final :

$$s_Y^2 = s_{y,x}^2 + s_r^2, \text{ avec}$$

$$s_{y,x}^2 = R^2 s_Y^2 \text{ et } s_r^2 = (1 - R^2) s_Y^2.$$

Le carré du coefficient de corrélation linéaire représente la variabilité de Y expliquée par une relation linéaire. La variabilité résiduelle définie au chapitre précédent à partir des résidus est inversement proportionnelle à R^2.

Si $r = 0$, toute la variabilité est résiduelle, les variations de X n'ont pas d'influence linéaire sur les variations de Y : la connaissance de X ne donne aucune information sur Y.

Si $r = 1$ ou $r = -1$, toute la variabilité est expliquée, et la relation entre Y et X est linéaire : la connaissance de X permet de prédire exactement la valeur de Y.

Reprenons la formule du test de la pente quand $\gamma = 0$, ie savoir si une pente observée est significativement différente de 0. Dans le cas où $\gamma = 0$, en remplacant la variabilité résiduelle par la nouvelle formule obtenue plus haut, on retrouve bien :

$$\frac{\hat{a}}{\sqrt{\frac{\sigma_r^2}{ns_X^2}}} = \frac{\frac{s_{XY}}{s_X^2}}{\sqrt{\frac{\frac{n}{n-2}(1-r^2)s_Y^2}{ns_X^2}}} \tag{41}$$

$$= \frac{s_{XY}}{s_X\sqrt{\frac{1}{n-2}(1-r^2)s_Y^2}} \tag{42}$$

$$= \frac{s_{XY}\sqrt{n-2}}{s_X s_Y\sqrt{1-r^2}} \tag{43}$$

$$= \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \tag{44}$$

qui est la formule donnée précedemment pour tester l'égalité de r à 0 dans le cas du test du coefficient de corrélation. On voit bien que si on a pas trouvé un r significativement différent de 0, on trouvera une pente égale à 0, ie pas de relation entre Y et X.

Si l'on veut mettre en parallèle la décomposition employée dans l'ANOVA1 et dans la régression, on peut l'écrire ainsi :

DIAPO 100 decomposition variance

ANOVA 1 $y_{ij} = \mu + a_i + e_{ij}$

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{p} n_i(\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

$$SCE_{tot} = SCE_{inter} + SCE_{intra} \quad \eta^2 = \frac{SCE_{inter}}{SCE_{tot}}$$

Modèle linéaire $y_{ij} = ax_i + b + e_{ij}$

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{p} n_i(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2$$

$$SCE_{tot} = SCE_{y,x} + SCE_{res} \quad R^2 = \frac{SCE_{y,x}}{SCE_{tot}}$$

Expliquer les variations de Y avec un modèle linéaire est moins général qu'avec une ANOVA 1, car dans le cas de l'ANOVA 1 on n'impose pas la contrainte que l'explication doit être linéaire.

$\Rightarrow$ La variabilité expliquée par le modèle linéaire est toujours inférieure à celle expliquée par l'ANOVA 1.

$\Rightarrow$ $R^2 \leq \eta^2$.

$\Rightarrow$ La quantité intéressante à étudier est la différence entre les moyennes de classes $\bar{y}_i$ et les estimations linéaires $\hat{y}_i$, qui sont les deux intermédiaires explicatifs.

Pour cela, on peut décomposer la variance inter-groupes de l'ANOVA1 à l'aide des $\hat{y}_i$, pour savoir quelle part de la variabilité expliquée par l'ANOVA est due à une explication linéaire :

$$\begin{aligned}\sum_{i=1}^{p} n_i(\bar{y}_i - \bar{y})^2 &= \sum_{i=1}^{p} n_i \left(\bar{y}_i - \hat{y}_i + \hat{y}_i - \bar{y}\right)^2 \\ &= \sum_{i=1}^{p} n_i(\bar{y}_i - \hat{y}_i)^2 + \sum_{i=1}^{p} n_i(\hat{y}_i - \bar{y})^2 \\ SCE_{inter} &= SCE_{Ecart} + SCE_{y,x}\end{aligned}$$

On a donc au final, en remettant les formules ensemble :

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{p} n_i(\bar{y}_i - \hat{y}_i)^2 + \sum_{i=1}^{p} n_i(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 \tag{45}$$

$$SCE_{tot} = \quad SCE_{Ecart} \quad + \quad SCE_{y,x} \quad + \quad SCE_{intra}. \tag{46}$$

Ou encore, en divisant tout par n et en remplacant chaque SCE par l'indice explicatif correspondant :

$$s_Y^2 = (\eta^2 - R^2)s_Y^2 + R^2 s_Y^2 + (1 - \eta^2)s_Y^2.$$

$\eta^2 - R^2$ est l'*indice de non-linéarité*. C'est la proportion de la variabilité expliquée par une ANOVA 1 et pas par la régression linéaire.

DIAPO 101 exemple graphique η et R.

$\eta^2 = R^2$ L'ANOVA 1 et la régression linéaire expliquent la même proportion de la variabilité globale : les données suivent bien un modèle linéaire.

$\eta^2 \gg R^2$ L'ANOVA 1 explique beaucoup plus de variabilité que la régression linéaire : le modèle linéaire ne correspond pas aux données, les variations de Y sont non-linéaires.

Attention, quel que soit le rapport entre η^2 et R^2, il ne faut pas oublier que si η^2 est très faible, on a globalement peut d'effet de X sur Y.

On calcule les CM en divisant les SCE par leur nombre de ddl associés. Les $p-1$ ddl de la variance *inter* dans le cas de l'ANOVA 1 se retrouvent décomposés en 1 ddl pour la régression (qui ne contient que 2 paramètres) et $p-2$ pour le reste. On a donc :

$$CM_{y,x} = \frac{SCE_{y,x}}{1} \quad CM_{Ecart} = \frac{SCE_{Ecart}}{p-2} \quad CM_{res} = \frac{SCE_{res}}{n-p}$$

Le test de linéarité se fait suite à une ANOVA1 pour laquelle on a rejeté H_0. Pour réaliser le test de linéarité, on va comparer les CM_{Ecart} et CM_{res} comme dans une ANOVA 1. Le test a la structure suivante :
H_0 : La relation entre Y et X, si elle existe, est linéaire.
H_1 : La relation entre Y et X, si elle existe, est non-linéaire.
On suppose par défaut une relation linéaire ; en effet le modèle linéaire est plus simple qu'un modèle polynomial ou exponentiel. On calcule ensuite :

$$F_{obs} = \frac{CM_{Ecart}}{CM_{res}},$$

et cette valeur est comparée à une valeur seuil à $(p-2, n-p)$ ddl, comme dans le cas de l'ANOVA 1.

Le test de linéarité ne peut pas être réalisé en dehors du cadre de l'ANOVA 1. La procédure globale de test doit être la suivante :

1. Vérifier que les hypothèses sont réunies, et réaliser une ANOVA 1.
2. Si l'ANOVA 1 est significative (effet du facteur), faire le test de linéarité pour savoir si l'effet est linéaire.

3. Sinon, s'arrêter : le facteur n'ayant pas d'effet, tester la linéarité de l'effet n'a pas de sens.

Variabilité	ddl	SCE	CM
Totale	$n-1$	SCE_{tot}	
Résiduelle (intra)	$n-p$	SCE_{res}	$CM_{res} = SCE_{res}/(n-p)$
Inter	$p-1$	SCE_{inter}	$CM_{inter} = SCE_{inter}/(p-1)$
Expliquée	1	$SCE_{y,x}$	
Ecart	$p-2$	SCE_{Ecart}	$CM_{Ecart} = SCE_{Ecart}/(p-2)$

Les 2 tests à effectuer à partir de ce tableau sont :
— Effet du facteur ? $F_{obs} = \frac{CM_{inter}}{CM_{res}}$, F_{seuil} à $(p-1, n-p)$ ddl.
— Linéarité de l'effet ? $F_{obs} = \frac{CM_{Ecart}}{CM_{res}}$, F_{seuil} à $(p-2, n-p)$ ddl.

DIAPOS 102-103-104 exemple final marmotte
DIAPO 105 pub BISM