UE11 : Biomédecine Quantitative
Jeudi 30 janvier 2014 à13h30
Matthieu Resche-Rigon

RT : Jérémie HUET
RF : Mehdi DAGHSEN

Cours numéro 1
Modélisation statistique

Le cours mis en ligne est très long : il fait 118 diapos. Je vais essayer de faire mon maximum pour mettre les informations importantes. Tout n'est pas à apprendre dans les diapos et ça le prof l'a répété plusieurs fois. Donc ne vous étonnez pas s'il y a des éléments dans les diapos non écrits dans la ronéo. Pour ne pas surcharger la ronéo, j'ai volontairement omis les diapos qu'il n'a pas présentées à l'oral. Les formules mathématiques ne sont, pour la plupart, pas à savoir, c'est de la « culture G ». En revanche le raisonnement est à connaître et il faut comprendre les principes derrière les modélisations.

SOMMAIRE

Introduction à la matière

I) Rappels

A) Définitions
- 1) Variables
- 2) Probabilité, Espérance, Variance
- 3) Opérations sur l'espérance et la variance
- 4) Médiane et quantile

B) Quelques lois

C) Statistique inférentielle et estimation

D) Loi Normale

II) Modélisation linéaire

A) Corrélation

B) Régression linéaire

III) Régression logistique

IV) Modèle multiple

A) Modèle linéaire multiple

B) Variable catégorielle non ordonnée

C) Modèle logistique multiple

Introduction à la matière

La biomédecine quantitative est une matière récente. Elle a pour objectifs de nous donner les outils statistiques nécessaires à la compréhension des études de recherche en médecine. Par ailleurs elle est sensée nous donner les bases nécessaires pour l'épreuve de LCA.

C'est un apprentissage progressif donc les cours de PACES sont supposés acquis (il y les polys de la fac sur didel ou les fiches médisup sur le weebly de l'an dernier). Les rappels dans ce cours sont les éléments les plus importants à savoir.

Il y aura 8 cours + 2 ED *(aucune info sur leur côté obligatoire ou non).* L'ED1 portera sur les cours 1, 2, 3 et l'ED2 sur les cours 4, 5, 6, 7, 8. Il y a en théorie à peu près 50 étudiants par groupe (*mais sous entendu en réalité beaucoup moins d'où le fait qu'ils soient plutôt non obligatoires).* Il faut aussi regarder l'emploi du temps en ligne car il y a eu quelques changements par rapport au poly de la fac. Bien sur le prof recommande vivement d'aller en ED : « très utiles pour la compréhension des cours ».

Tous les cours seront disponibles sur didel.

En ce qui concerne les contrôles continus, il y aura un contrôle continu blanc d'ici peu et deux sessions de controles continus : le premier sur les cours 1, 2, 3 et le deuxième sur les cours 4, 5, 6. Le prof ne sait pas encore si ça va être en distanciel comme l'année dernière ou si ça va être comme pour la gastro : infos à venir ! Dans tous les cas il faudra se connecter sur didel.

Coefficients note finale : 20% CC + 40% QCM + 40% QR (rattrapages : 50%QCM + 50% QR).

I) Rappels

Tous les modèles sont faux ! La seule question qu'il faut se poser c'est : « est-ce que c'est une bonne approximation de la réalité ? ».

A) Définitions

1) Variables

Une **variable aléatoire** est un phénomène aléatoire X auquel on s'intéresse dans une population. On ne peut pas le prédire même si l'on peut avoir une idée de là où il se trouve. Par exemple si on regarde le poids des individus dans une population, on ne va pas pouvoir prédire le résultat mais en fonction de la taille, on va pouvoir s'attendre à un intervalle dans lequel va se situer le poids.

On distingue deux grands types de variables :

• les **variables (ou données) qualitatives** : elles sont non mesurables sur une échelle. Elles sont elles-mêmes divisées en :

- **données ordinales**, c'est-à-dire avec relation d'ordre (exemple : stades d'un cancer I, II, III, IV)

- **données catégoriques**, donc sans relation d'ordre (exemple : homme/femme, brun/roux/blond/chauve…) (*c'est là où le prof a failli dire blanc/noir/jaune/marron mais s'est retenu ☺)*

• **les variables (ou données quantitatives)** : elles sont mesurables sur une échelle pour le coup. Divisées en 2 sous groupes :

- **données continues** : des valeurs réelles. Elles sont en fait considérées comme rélles car elles sont en réalité limitées par la précision des instruments de mesure. Exemple : poids, taille, âge, glycémie.

- **données discrètes** : des valeurs isolées ou plutôt par intervalles de valeur. Exemples : nombre d'enfants, de métastases, âge au dernier anniversaire…

Remarque : certaines variables peuvent changer de type en fonction de la méthode de mesure qu'on utilise. Par exemple pour l'âge, on peut choisir un classement vieux/jeune (donnée ordinale ou catégorique selon comment on pose la question), l'âge exacte (donnée continue) ou l'âge au dernier anniversaire (donnée discrète).

Attention : la variable X désigne la grandeur mesurée (exemple : le poids) tandis que la valeur t (ou x) désigne le résultat obtenu (exemple : 67,5 kg) !

2) Probabilité, Espérances, Variances

La probabilité de t associée à une variable X correspond au nombre de fois qu'on va tomber sur l'intervalle contenant t sur le nombre de tirages au total, si on prend un nombre de tirages infini.

Ainsi pour une **variable discrète**, l'intervalle contenant t est fini donc P(t) est défini et donc pour toute valeur t, **$P(t)=P(X=t)$.**

La fonction de répartition est la fonction qui à toute valeur associe la probabilité que le tirage se trouve en dessous de cette valeur. En écriture mathématique, ça donne : $F(t)=P(X\leq t)$. Ainsi pour une variable discrète **$F(t)=\Sigma P(X=t_i)$** avec t_i représentant chaque valeur inférieure ou égale à t.

L'Espérance est ce qui correspond à la moyenne. C'est la somme des valeurs obtenues divisée par le nombre de tirages, si on faisait un nombre de tirages infinis. Donc pour une valeure discrète, **$E(X)=\mu=\Sigma t_i P(X=t_i)$.**

La Variance est la moyenne de la somme des carrés des écarts à la moyenne. $Var(X)=\sigma^2$. Pour une valeur discrète **$Var(X)=\Sigma(t_i-\mu)^2P(X=t_i)$.**

Pour les **valeurs continues**, le principe est le même mais les formules sont différentes. En effet, comme l'intervalle contenant t est considéré comme infiniment petit, alors la probabilité associée à une valeur t est nulle car il y une infinité d'intervalles. Ainsi on utilise la densité et non plus la probabilité. Si on prenait une distance h suffisamment petite pour que la densité ne varie pas sur cette distance, alors on aurait :

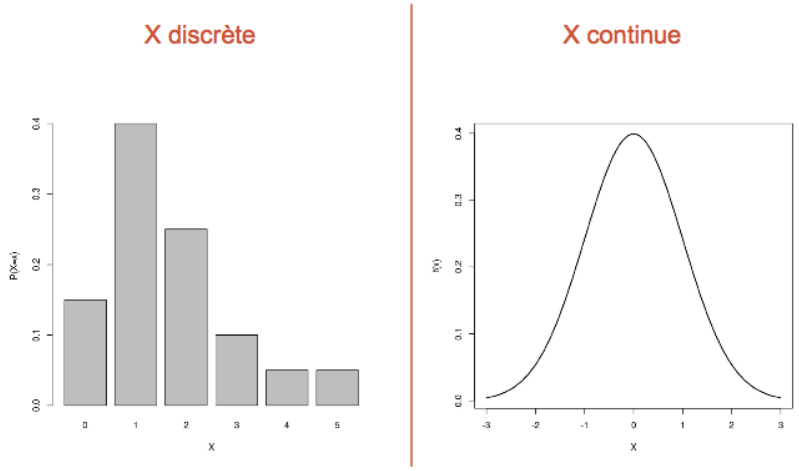
$P(t\leq X\leq t+h)=h*$densité de X au point t.

Donc la densité **f(t)=limite quand h tend vers 0 de $P(t\leq x\leq t+h)/h$.**

De même la fonction de répartition en fonction de b d'une valeur continue est la somme des densités multipliées par un petit dt sur lequel f(t) ne varie pas entre -∞ et b. Comme cette somme est une somme infiniment grande d'éléments infiniments petits, on n'utilise plus un Σ mais une intégrale, d'où $F(b)=P(X\leq b)=\int_{-\infty}^{x} f(t)dt$

$$E(X)=\mu=\int_{-\infty}^{+\infty} tf(t)dt \quad \text{et} \quad Var(X)=\sigma^2=\int_{-\infty}^{+\infty}(t-\mu)^2 f(t)dt$$

Probabilités et densités



3) Opérations sur l'espérance et la variance

•**Espérance** : Soient X et Y deux variables aléatoires, a et b deux réels :
E(aX+bY)=aE(X)+bE(Y)

•**Variance** : Elle permet de caractériser la dispersion d'une distribution et c'est très important pour savoir dans quelle mesure tel traitement marche par exemple.

Var(aX)=a²Var(X)
Var(X+Y)=Var(X)+Var(Y)+2Cov(X,Y)
Var(X-Y)=Var(X)+Var(Y)-2Cov(X,Y)
Var(E)=E[(X-E(X))²]=E(X²)-E(X)²
Cov(X,Y)=E(X*Y)-E(X)*E(Y)

La covariance est nulle pour deux variables indépendantes. Mais si on s'intéresse par exemple à la pression artérielle différentielle (PAS-PAD), on voit bien que la PAS et la PAD sont liées déjà par le fait que PAD est toujours inférieure à la PAS. Donc la variance de la différence de ces deux variables aléatoires va être surestimée si on ne soustrait pas la covariance. L'important est de comprendre la signification de cette covariance, on ne nous demandera jamais d'en calculer une.

4) Médiane et quantile

La médiane est la valeur pour laquelle la fonction de répartition est égale à un demi : P(X≤Médiane)=0,5.

De même un percentile est une valeur pour laquelle P(X≤perc)=p.

p usuels : 2.5ème, 5, 10, 25 (1er quartile), 50 (médiane), 75 (3ème quartile), 95, 97.5ème

Par exemple, pour les courbes poids/taille sur le carnet de naissance, on se sert des 2,5ème et 97,5ème percentile pour savoir si on sort de la normalité.

/!\Il faut comprendre que ça ne signe pas une pathologie mais que c'est uniquement un signal d'alerte. (Canguilhem : anormal n'est pas pathologique)

B) Quelques lois

Lois discrètes

• **Bernoulli** : loi en oui/non d'où X est égale à 0 ou 1 et P(X=1)=p.

E(X)=p et Var(X)=p(1-p)

•**Loi Binomiale** : on répète n fois une expérience de Bernoulli. D'où X=0, 1, …, n. X suit une loi binomiale de paramètres n et p : bin(n,p).

E(X)=np et Var(X)=np(1-p)

•**Loi de Poisson** : utilisée pour estimer un nombre de quelque chose par unité de temps, surface ou volume. Par exemple le nombre d'individus qui se présentent à Bichat par heure.

Elle dépend d'un seul paramètre λ, et **E(X)=Var(X)=λ**

Lois continues

•**Loi Uniforme** : X appartient à [a ;b] et **E(X)=(a+b)/2**

1 seule utilité : tirage au sort

•**Loi exponentielle** : très utilisée par exemple pour des décès au cours du temps ou pour la désintégration radioactive car c'est une loi sans mémoire.

Elle varie en exponentielle d'un paramètre λ.

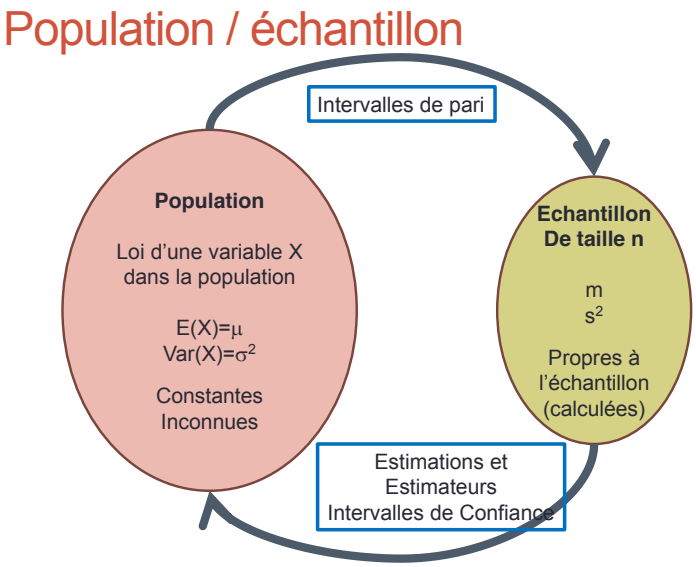
E(X)=1/λ et Var(X)=1/λ²

•**Loi de Gauss ou loi Normale** : très très utilisée, surout en médecine.

Elle a deux paramètres : **μ et σ²** qui sont respectivement égaux à l'espérance et à la variance.

C) Statistiques inférentielles et estimation

Une statistique est une fonction d'observation qui essaie de décrire un paramètre de population. Par exemple on va essayer d'estimer la moyenne d'âge de la population étudiante en observant la moyenne d'âge dans un amphi.



Il y aura donc la **variable vraie** de la **population** et **l'estimation des paramètres** de cette variable dans un **échantillon**. Les lettres grecques sont utilisées pour les variables vraies et les lettres normales pour les estimations. Mais les calculs de moyenne ou de variance restent identiques.

On va donc pouvoir à partir d'une variable vraie connue faire des **intervalles de pari** pour un échantillon de taille n. C'est-à-dire que par exemple on va se dire : il y a 95% de chances que ma moyenne mesurée se trouve dans tel intervalle.

Inversement, et c'est beaucoup plus utile, on va pouvoir estimer des paramètres de la variable vraie à partir de données mesurées. On va donc définir des **intervalles de confiance** (IC): « je pense que la vraie valeur se situe à 95% de chance dans tel intervalle ».

La définition à notre niveau d'un intervalle de confiance à x% (même si fausse) est donc : il y a x% de chances que la vraie valeur s'y trouve. On utilise généralement l'IC à 95%.

La plupart des IC sont sous la forme d'un intervalle $T \pm \Delta$ où T est l'estimation ponctuelle et Δ est l'inverse de la précision de l'estimation et dépend de n.

Or la variance varie en $1/n$ mais c'est du carré donc la précision varie en $1/\sqrt{n}$. Ainsi si on multiplie n par 100, on augmente la précision par 10 seulement.

D) Loi Normale

Les valeurs de f(t) et F(t) ne sont pas à connaître.

Si X suit une loi normale de paramètres μ et σ alors $Z=(X-\mu)/\sigma$ suit une loi N(0,1). On va donc, par simplification, toujours se référer à cette loi N.

Propriétés de la loi Normale :

- $F(\mu)=0,5$
- 95% des valeurs sont dans l'intervalle $[-2\sigma;+2\sigma]$

Théorème Limite Central : Soit X une variable aléatoire d'espérance $E(X)=\mu$ et de variance $Var(X)=\sigma^2$. Pour un échantillon de taille n, on a $M_n=\Sigma t_i/n$.

La loi de M_n converge asymptotiquement quand n tend vers l'infini vers la loi normale de paramètres μ et σ^2/n.

Donc pour n assez grand **on peut approximer M_n par une loi normale**.

Dans ce cas là, on a l'intervalle de confiance à 1-α de la moyenne μ de X est donc : $M_n \pm \varepsilon_\alpha(\sigma/\sqrt{n})$.

Cas particulier : $\mathbf{IC_{95\%} \approx m \pm 2(s/\sqrt{n})}$

II) Modélisation Linéaire

La modélisation linéaire a pour but d'analyser la **liason entre deux variables**. Ça arrive tout le temps en médecine ! Entre poids et taille, entre quantité de médicaments et effets, etc…

Pendant toute la suite des explications, on va donc considérer deux variables aléatoires X et Y que l'on considère à priori comme indépendantes donc non liées= **hypothèse H_0**. On va donc essayer de rejeter H_0. Mais attention ce n'est pas parce qu'elle ne sont pas liées qu'on ne va pas associer deux mesures. Par exemple à tous les individus qui font 70kg on va associer leur taille.

On va donc faire des séries de **doubles mesures a priori indépendantes** : (X_1,Y_1) ; (X_2,Y_2) ; … ; (X_n,Y_n).

Il y a donc deux situations en fonction de la question qu'on pose. Les deux peuvent paraître identiques mais en statistiques la question posée doit être très précise et est très importante pour l'interprétation (LCA !!) :

- On recherche une **symétrie entre X et Y** : on veut juste savoir s'il y a une corrélation entre ces deux valeurs à priori indépendantes.
- On recherche une **asymétrie entre X et Y**, c'est-à-dire qu'on va essayer de prédire l'un en fonction de l'autre. On va utiliser pour cela une régression linéaire (car on est dans une modélisation linéaire). On met donc en avant une valeur plutôt qu'une autre.

A) Corrélation

L'objectif va être de calculer un **coefficient de corrélation** (soit de Pearson ou de Spearman) et en fonction de sa valeur on va pouvoir démontrer une plus ou moins forte corrélation. Le problème c'est qu'on est en statistique donc il n'y a pas de valeur ponctuelle mais plutôt un intervalle de confiance autour d'une valeur centrée.

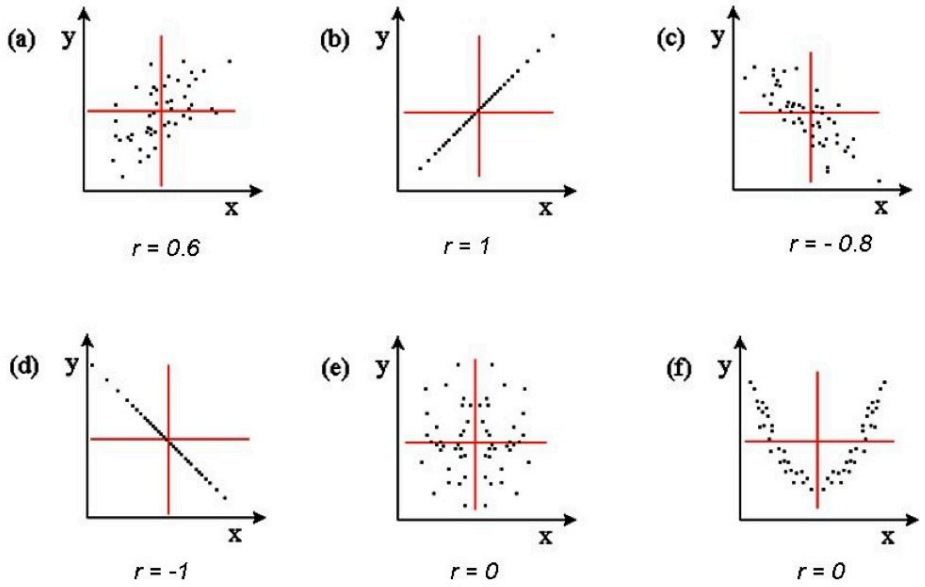
On va donc définir **ρ qui correspond à la valeur vraie du coefficient de Pearson** et r sa valeur mesurée dans un échantillon, son estimation. r n'a donc aucune utilité tout seul !! Il doit toujours se présenter dans un IC !! *(s'il y a un truc à retenir dans ce cours c'est cette nuance).*

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad \text{et} \quad r = \frac{\sum (x_i - m_X)(y_i - m_Y)}{\sqrt{\sum (x_i - m_X)^2 \sum (y_i - m_Y)^2}} = \frac{s_{XY}}{s_X s_Y}$$

ρ et r sont compris entre -1 et 1 par définition.

- ρ=1 ou ρ=-1 indique une relation linéaire parfaite entre X et Y : le tracé de Y en fonction de X donne des points alignés.
- ρ=0 : pas d'association **linéaire**.

Donc plus ρ^2 se rapproche de 1, plus la corrélation est forte. Ainsi on dit que ρ est la proportion de variance de Y qui s'explique par la variabilité de X.



Sur les différents graphes ci-dessus on a des représentations des y_i en fonction des x_i (valeurs mesurées). On remarque que pour **r négatif la pente de la droite est négative et inversement.**

On remarque aussi que r=0 dans le cas (f) mais pourtant on voit bien une corrélation apparente entre les deux. En fait il s'agirait plutôt ici d'une corrélation hyperbolique (l'effet augemente aux extrêmes) qu'une corrélation linéaire.

Donc un ρ nul ne veut pas dire qu'il n'y a pas d'association !!!

Lorsqu'on aura calculé notre r, **on va le tester par rapport à l'hypothèse H_0**. On va donc créer un intervalle de confiance autour de r et vérifier que cet intervalle ne contienne pas 0. Donc contrairement à ce qu'on a appris en PACES, **on va pouvoir raisonner sur un IC**. On va généralement choisir un IC à 95%. La méthode de calcul de cet IC n'est pas importante (*mais il est sur les diapos si ça vous intéresse).*

/!\ Il ne faut pas confondre le test de ρ à 0 et une corrélation forte ! Avec un nombre suffisamment grand de sujets, on peut montrer une corrélation à 0,1 mais c'est pas ça qui va la rendre significative.

Corrélation	Positive	négative
Parfaite	entre 0.98 et 1	entre -1 et -0.98
Forte	entre 0.80 et 0.98	entre -0.98 et -0.80
Moyenne	entre 0.60 et 80	entre -0.80 et -0.60
Faible	entre 0.35 et 0.60	entre -0.60 et -0.35
Négligeable ou nulle	entre -0.35 et 0.35	

B) Régression linéaire

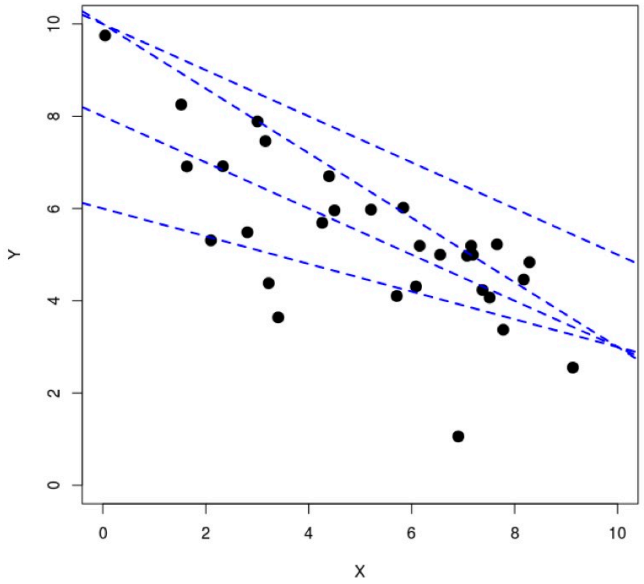
Y est appelée variable dépendante ou expliquée et X variable indépendante ou explicative.

L'objectif ici est de pouvoir prédire Y en fonction de X donc donner une **fonction linéaire de l'espérance de Y sachant X** sous la forme **$E(Y/X)=\alpha+\beta X$**. On cherche donc à estimer E(Y/X=x) sous la forme d'une fonction de X. Mais on n'a pas $Y=\alpha+\beta X$!!

En effet X et Y sont toutes les deux des variables aléatoires, il n'y a donc **pas de relation déterministe entre Y et X** (ce n'et pas parce qu'on fait 1m70 qu'on va forcément peser 70kg !). À une valeur de X correspond plusieurs valeurs de Y.

On peut alors écrire **$Y=\alpha+\beta X+\varepsilon$** où ε est une variable aléatoire d'espérance nulle qui représente l'ereur résiduelle du modèle. En pratique on suppose que ε suit une loi normale $N(0,\sigma^2)$.

Il faut maintenant déterminer α et β. Quand on a un graphe de tous les couples (x_i,y_i), on va devoir tracer une droite de régression linéaire mais laquelle choisir ?



Quelle courbe en pointillées choisir ?

On va utiliser la **méthodes des moindres carrés**, c'est-à-dire que l'on va choisir la droite pour laquelle la somme du carré des distances des points à la droite est minimale.
L'ordinateur va donc nous donner des valeurs a et b qui minimisent cette somme.

a correspond à l'ordonnée à l'origine ou intercept et b est la pente de la droite ou l'effet de X sur Y.

Comme la méthode du choix de a et b est arbitraire et n'est pas vraie, on va également définir un IC à 95% de ces deux valeurs dans lequel on trouvera (à 95% de chances) les vraies valeurs α et β.

Cet IC à 95% peut s'obtenir à partir de l'estimation ponctuelle de la variance grâce à la loi de Student.

En pratique dès que **n est grand**, on peut approximer la loi de student par une loi normale : **$IC_{95\%}\approx b\pm 1{,}96 s_b$**

Rappel des hypothèses :
- Indépendance des observations (x_i,y_i)

- Linéarité de relation entre Y et X
- Normalité des ε.
- X et Y suivent des lois continues sur]-∞ ; +∞[!!!

III) Régression logistique

Maintenant qu'on a fait ça, le problème c'est que Y est très souvent une variable discrète binaire en 0 ou 1. Par exemple on va regarder les Infarctus Du Myocarde (IDM) en fonction du poids ou le cancer en fonction de la consommation de tabac. C'est même le type d'étude le plus courant. **D'un côté on aura une variable continue et de l'autre une variable discrète**, ce qui n'est pas du tout comparable en mathématiques (c'est comme essayer de comparer des pommes et des oranges).

On va donc essayer de transformer le problème pour se ramener au problème de départ qu'on sait résoudre *(c'est à peu près toujours comme ça en maths ☺)*. L'objectif est donc de transformer un espace ne contenant que les éléments 0 et 1 en un espace contenant tous les réels donc R *(eh oui c'est possible de transformer des pommes en oranges en maths).* Pour cela il y a trois étapes :

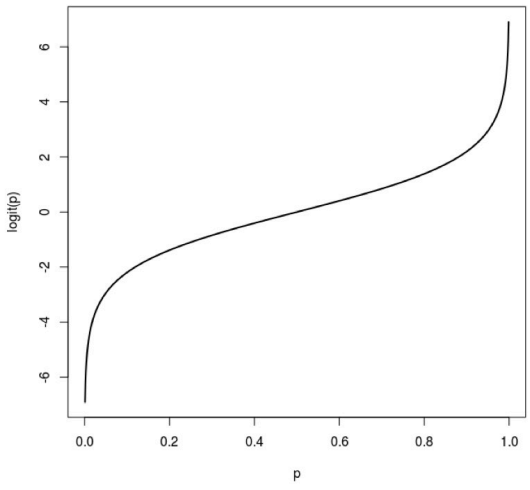
- Premièrement au lieu de prendre des valeurs de y_i donc 0 ou 1 on va choisir **l'espérance de Y** associé à chaque x_i, qui correspond à la **probabilité p** de Y. Or p est compris entre 0 et 1. Donc on est passé d'un espace à deux valeurs à un intervalle [0 ;1].
- Ensuite on va **diviser p par (1-p)**. On passe donc de [0 ;1] à [0 ;+∞[car 0/1=0 et 1/0=+∞
- Enfin on va prendre le **logarithme néperien de tout ça**. On passe donc de [0 ;+∞[à]-∞ ;+∞[car ln(0)=-∞, ln(1)=0 et ln(+∞)=+∞ et ln est une fonction croissante.

Pourquoi avoir choisi une telle seérie de fonctions ? Parce que ses propriétés nous arrangent bien en plus, explications plus bas.

On définit donc la fonction **logit(p)=ln(p/1-p)**. Sa **fonction inverse $logit^{-1}(X)=f(X)$ est la fonction logistique**. Elle est telle que f•logit(x)=1. Donc avec une valeur de logit(p)=c on va pouvoir retrouver p en appliquant la fonction f à c.

On a f(X)=[exp(X)]/[1+exp(X)]

Voici à quoi ressemble logit(p) :



On va ensuite faire une régression logistique simple comme pour la modélisation linéaire : on aura **logit(p)=α+βX.**

L'avantage de cette fonction est la simplicité du calcul de l'odds ratio (OR). En effet, l'odds ratio entre deux valeurs de x : x_0 et x_1 est égale à $(p_1/1\text{-}p_1)/(p_0/1\text{-}p_0)$.

Or **$\text{logit}(p_1)-\text{logit}(p_0)=\log(OR)=\beta X$**.

Donc chaque fois qu'on augmente d'une unité X, on augmente l'O.R de β. **C'est la log-linéarité de l'OR**. Or ce qui nous intéresse la plupart du temps c'est le sur-risque, c'est pas le risque absolu en lui-même car on n'a pas tous le même risque de base. Donc cette log linéarité est très intéressante. D'où le choix de la fonction logit.

Pour estimer α et β, on utilise des fonctions assez complexes qui ne sont pas à savoir.

$$L(a,b)=\prod_{i=1}^{n}\left(\frac{exp(a+bx_i)}{1+exp(a+bx_i)}\right)^{y_i}\times\left(\frac{1}{1+exp(a+bx_i)}\right)^{1-y_i}$$

On va chercher des **valeurs de a et b qui maximisent la vraisemblance L** et pour cela on utilise l'estimateur du maximum de vraisemblance. Elles sont obtenues par des algorithmes de maximisation (Newton Raphson).

IV) Modèle multiple

A) Modèle linéaire multiple

Il arrrive que Y ne dépende pas que d'une seule valeur X mais **d'un ensemble de valeurs** (X_1, X_2, …, X_k) : exemple de l'IDM qui va dépendre de l'alcool, du cholestérol, du tabac, etc… On écrit alors un modèle linéaire multiple :

$Y=\beta_0+\beta_1X_1+\ldots+\beta_kX_k$

On va donc **généraliser le modèle de régression** simple à un modèle multiple. On **veut tester chaque β_i à 0**. L'estimation des paramètres se fait aussi par minimisation de la distance entre les valeurs prédites par le modèle et les observations (méthode des moindres carrés). On obtient alors un vecteur de paramètres (b_0,b_1,…, b_n).

Chaque b_i pour n assez grand suit une loi normale. Donc on pourra facilement créer des IC également et tester chaque facteur à 0.

On dit que **l'effet de X_i est ajusté sur les autres variables**. C 'est ainsi qu'on a pu démontrer que l'effet de l'alcool était indépendant du tabac pour l'IDM.

B) Variable catégorielle non ordonnée

Variable	Dummies		
GS	X_1	X_2	X_3
O	0	0	0
A	1	0	0
B	0	1	0
AB	0	0	1

On va créer des **dummy variables**. Si la variable possède n catégories, on crée n-1 variables. Par exemple, pour le groupe sanguin (ci-dessus) il y a trois variables. Pour chaque groupe sanguin on va mettre une variable à 1 et le reste à 0. Il y aura également une variable avec que des 0 (ici le groupe O).

Donc **les coefficients qu'on va estimer correspondent à ceux des « dummy » variables.** Ils représentent le log(OR) par rapport à une catégorie de référence pour laquelle toutes les variables valent 0 (ici le groupe sanguin O).

Les tests sont des **tests de Wald partiels** et ne concernent que chaque dummy variable.

C) Modèle logistique multiple *(le prof est passé très vite dessus)*

C'est la même chose que pour la régression logistique unique mais **généralisé à plusieurs variables explicatives X_i.**

On a donc **$P(Y=1)=\exp(\Sigma\beta_i X_i)/[1+\exp(\Sigma\beta_i Xi)]$** avec $\Sigma\beta_i Xi= \beta_0+\beta_1 X_1+...+\beta_j X_j$

$\Sigma\beta_i Xi$ est appelé le prédicteur linéaire.

Pour la formule de la vraisemblance on généralise également à $\Sigma\beta_i Xi$.

Néanmoins il y a une règle supplémentaire : le nombre de Xi ne doit pas dépasser 1 pour 10 évènements ou non évènements. C'est la **règle du pouce**. Par exemple si on reprend l'exemple de l'IDM avec l'alcool et le tabac, on devra avoir au moins 20 sujets ayant eu un IDM.

DÉDICAAAAACE !!!!

À ÉBISOL !!!! bien sûr parce que ça reste l'asso la plus stylée de la fac !! ☺ dont :

- Éléonore : pour toutes tes escapades en cours pour aller te remaquiller, pour toute ta joie de vivre légendaire, pour ta proximité géographique, pour tous nos voyages en métro, pour ton accent anglais : gros gros love my bunny rabbit, (et encore merci pour votre album)
- Nora : parce que je sais que tu vas regarder avec attention ma dédicace : pour tes golden standards (maud t'es une voleuse), pour ma nouvelle irlandaise préférée, pour ton humour déjanté, pour ton énergie inépuisable, pour notre amour pour la musique (la vraie ☺)
-Clara : pour ta droiteté légendaire, pour ton engagement et surtout voy a cerrar el agua. Finalement on a les mêmes défauts mais on se supporte pas trop mal
- Pauline : pour avoir payé ta maison tous les weekends, pour ton sourire, pour ta gentillesse, pour ta fidélité et snoop dogg
- Maud : gros coup de cœur de l'année, je sais pas ce qu'on ferait sans toi à EBISOL, merci énormément
- Audrey : pour avoir tenu ta promesse au long de l'année, vive la femme ! Continue à t'imposer dans ce monde de brutes
- Aux mecs d'ébisol : Antoine , finalement la faluche c'est pas si mal que ça, Gauthier, toujours bien entouré, Gwen, pour ta prestation à notre viéo de présentation
-Éléonore : co p1, et maintenant dans la même asso… quel voyage ! dsl les gouters solidaires j'ai pas trop le temps d'y aller en ce moment
-au pole A/P : Marine je savais que je pouvais compter sur toi, j'ai bien fait de t'embarquer là dedans, Alice gros big up aussi
-A lorraine, laurène, louise, laure, Océane, Julie grandes découvertes aussi de cette année !!

A Emma parce que j'ai cité les deux autres du Pérou, bien sur j'allais pas t'oublier. Mais qu'est-ce que t'es allée foutre au bda ?

A Maxime, si tu lis cette ronéo avant le mois de mai…

A alessandra, il est où mon portable ??!

Au gleechat : géniales comme initiatives Lydie, Laura oui ça me soule parfois de dire bonjour à tout le monde le matin mais jamais toi, toi c'est parce que je te vois pas (je suis aussi fatigué), à Paul, à Solène, que de nouvelles rencontres également : il faut controler sa respiration !

A mes deux fillottes : chloé et marion, à ma marraine Sarah en D2 : courage !

A tous les D2 que je connais, aux anciens d'ebisol (Fanny, Karim, Elise, Marine, Benoit) je vous oublierai jamais !

A Stan : eh ouais j'ai pris ma revanche en P1. Tu ne me battras plus jamais en physique ☺

Aux assos de la fac ! A l'avc : qui aurait cru qu'une petite bande de potes en p1 à laennec arriverait à accomplir de si grandes choses, je suis fier de vous, j'aime cette fac ! Au BDA, au BDS (vive le basket !), au Tuto.

A Sheldon ou Jim Parsons : big Fan !! Bazinga. I'll be one of your C-men if you want

Enfin à tous ceux que j'aurais oublié : il est un peu tard et je m'excuse d'avance pour ces oublis ! Je tiens à signaler que l'ordre ne reflète aucune différence d'appréciation.

Vive le québec libre !

LE 14 FEVRIER ON N'OUBLIE PAS DE PASSER A LA FRANCHIZ POUR LA SAINT VALENTIN ET ON DÉPOSE SON MOT D'AMOUR AVANT DANS LA SEMAINE.

LE 7 MARS JOURNÉE DE DON DU SANG ON PRÉPARE SES GLOBULES ROUGES

À Nina <3