Problèmes de limites en statistique

Joseph SAINT PIERRE,
Centre Interuniversitaire
de Calcul de Toulouse
118, Route de Narbonne
31062 Toulouse Cedex 04
Tél: 05 61 36 60 21
Fax: 05 61 52 14 58
e-mail: Joseph.Saint-Pierre@cict.fr

Vendredi 24 Mai 2002

Résumé

Un des problèmes les plus fréquents en statistiques consiste à extrapoler des résultats et de dépasser des limites inaccessibles. Certains pourcentages ne peuvent pas descendre en dessous de 0 ou passer au dessus de 100, on ne pourra jamais courir un cent mètres en moins de 0 seconde. Le but de cet exposé est de donner quelques idées pour aborder ce type de situations pour des chercheurs qui ne sont pas spécialistes des statistiques mathématiques.

1 Mise en garde

Attention cet exposé est une mise en garde contre des extrapolations abusives dans la modélisation statistiques, mais il n'existe pas une méthode mathématique pour faire face à ce problème et encore moins une méthode disponible dans les logiciels de statistiques permettant de résoudre simplement les difficultés de ce style. Cet exposé a été imaginé dans le cadre d'une réflexion beaucoup plus large sur les dangers d'une utilisation inconsidérée des statistiques et surtout pour mettre en garde les utilisateurs de logiciels de statistiques, face à la facilité apparente des calculs. Le problème de l'extrapolation de tendances linéaires locales à des situations beaucoup plus globales est un des dangers de la statistique qui me paraît le plus intéressant mais il n'est pas le seul. Cet exposé se présente plus comme une prévention que comme une initiation à l'utilisation de méthodes. Il existe par ailleurs dans l'université des formations à l'utilisation des méthodes, dont certaines sont utilisées ici et des formations à l'utilisation des logiciels. Par exemple la régression linéaire est présentée dans cet exposé mais il ne s'agit pas d'expliquer les principes de la régression linéaire ni d'expliquer comment on lance une régression linéaire avec tel ou tel logiciel.

2 Préambule

Les probabilités et les statistiques peuvent concerner les sujets les plus ordinaires de notre vie quotidienne, par exemple : les tartines beurrées ou couvertes de confiture tombent elles plus souvent du mauvais côté ? Il est commun de penser que les tartines atterrissent toujours du côté beurré ou confituré et cela est connu comme étant un des corollaires de la fameuse loi de Murphy, lorsqu'une chose peut aller mal, elle va mal... Le physicien (britannique bien sûr) Robert Matthews de l'université d'Aston a tenté de donner une explication rationnelle à ce fâcheux phénomène, en répétant plusieurs fois l'expérience, donc en faisant des statistiques, on peut constater qu'un livre qui tombe du rebord d'une table, acquiert un mouvement de rotation et compte tenu de la hauteur des tables et de la taille habituelle d'un livre, celui ci a le temps de faire un demi-tour avant d'atteindre le sol, on peut extrapoler le cas des livres à celui des tartines, c'est du moins ce qu'a fait Robert Matthews. Il semble donc, d'après ces observations que ce phénomène désagréable ne serait pas dû au pur hasard ni même à la loi de Murphy. La hauteur des tables est déterminée par la taille des êtres humains et celle ci n'est pas aussi variable que certains pourraient le croire.

Il se trouve que le terme régression, extrêmement employé en statistiques et qui désigne un ensemble de méthodes (régression linéaire, multiple, logistique etc.), a une origine liée à la taille des êtres humains. Ce terme est dû à Francis Galton (1822-1911), un cousin de Charles Darwin, anthropologue et explorateur. Galton s'est beaucoup intéressé à l'hérédité de l'intelligence, il a été un des premiers à utiliser des questionnaires et des méthodes d'enquêtes, il était un partisan des théories héréditaristes mais il s'est aussi intéressé à l'anthropologie physique. En 1889 il a étudié la relation entre la taille des parents et des enfants et il a remarqué les faits suivants, les parents de petite taille ont des enfants qui sont aussi de petite taille mais qui sont légèrement plus grands et les parents de grande taille ont des enfants qui sont aussi de grande taille mais qui sont légèrement plus petits. Il s'agit bien sûr de tendances générales et non de faits systématiques, Galton a été surpris par le fait que les enfants de parents de grande taille soient souvent plus petits que leurs parents et il considéra que ces enfants régressaient, dans le sens ordinaire du terme . Galton utilisa l'expression de régression vers la moyenne.

Il semble bien que la taille des être humains soit limitée et qu'il est impossible par sélection eugéniste d'avoir des individus dépassant une taille maximale qui n'est pas exactement connue mais qui est bien inférieure à trois mètres. Compte tenu de la gravité de la Terre et des caractéristiques biomécaniques des mammifères, la position debout est incompatible avec une taille trop grande, le crâne deviendrait beaucoup trop facilement cassable lors d'une chute de trois mètres.

La taille des humains peut évoluer dans le temps, pour un sujet donné au cours de sa vie, elle augmente de la naissance jusque vers l'adolescence, pour des populations elle peut augmenter en raison de meilleures conditions alimentaires mais ces augmentations sont nécessairement limitées dans le temps. Des augmentions de la taille moyenne de plusieurs centimètres en quelques décennies pour un pays se sont parfois produites suite à de profondes modifications du mode de vie, par exemple au Japon après la deuxième guerre mondiale. De telles augmentations si elles se prolongeaient quelques siècles donneraient des tailles sans limite. Bien évidemment cela est impossible, il existe une limite, non précisément connue mais qui constitue une barrière infranchissable.

3 Taille d'adolescents, un cas continu

À partir d'une étude réelle, il est possible de percevoir, pour la croissance de la taille humaine, en quoi il est absurde d'extrapoler certains résultats à partir d'une régression linéaire simple. 208 élèves du collège de Lalande à Toulouse ainsi que 64 élèves de cours moyen deuxième année, supposés passer en sixième dans ce collège, ont rempli un questionnaire élaboré par les élèves du collège afin de s'initier aux statistiques. Parmi les questions posées il y avait l'âge, ou plutôt l'année de naissance, ce qui revient à peu près au même, et la taille. Compte tenu des informations disponibles par les enseignants, et des estimations visuelles il semble que les réponses sur ces deux questions soient correctes mais il y a un certain nombre de réponses manquantes. Les élèves ont entre 10 et 17 ans. Le collège se situe dans une ZEP et beaucoup d'élèves sont en retard.

Voici les résultats, obtenus avec SPSS, de la régression linéaire de la variable taille sur la variable âge. Aucune autre variable n'a été prise en compte, les filles et les garçons sont mélangés alors que la croissance de la taille est différente suivant le sexe dans les âges considérés et cela est très perceptible dans cette population.

```
Dependent variable.. TAILLE            Method.. LINEAR


Multiple R           .73689
R Square             .54301
Adjusted R Square    .54111
Standard Error      8.20605

            Analysis of Variance:

                DF   Sum of Squares      Mean Square

Regression       1        19283.517        19283.517
Residuals      241        16228.780           67.339

F =     286.36333       Signif F =  .0000

Variable                   B        SE B       Beta         T  Sig T

AGE                 4.944773     .292205    .736892    16.922  .0000
(Constant)         88.330923    4.006557               22.047  .0000
```

L'interprétation des résultats est évidente entre l'âge de 10 ans et l'âge de 17 ans la taille dépend de l'âge, les élèves les plus âgés sont les plus grands. La pente de la régression vaut 4,94 soit près de 5, cela veut dire qu'en moyenne il y a 5 centimètres d'écart en taille par année, un élève de 15 ans mesure environ 15 centimètres de plus qu'un élève de 12 ans. Il y a bien sûr une variation résiduelle, non expliquée. La statistique F de Fisher est hautement significative et l'âge a un effet incontestable sur la taille. Cela n'est certainement pas une découverte... Mais l'utilisation du modèle de régression ne peut certainement pas servir à prédire la taille d'un sujet d'un âge supérieur à 17 ans ou même inférieur à 10 ans. En effet la taille à la naissance prédite par ce modèle est tout simplement la valeur de la constante c'est à dire 88,33 et cela voudrait dire que la taille moyenne des enfants à la naissance est supérieure à 88 centimètres ce qui est une surestimation grossière et totalement irréaliste. La taille moyenne des enfants à la naissance étant de l'ordre de 50 centimètres cela veut dire que la croissance de la taille dans la petite enfance est encore plus importante que celle observée durant l'adolescence, cela aurait besoin d'être étudié plus finement. Par ailleurs il

y a des périodes de croissance très rapides à l'adolescence mais l'âge où la croissance est la plus importante est très variable d'un sujet à un autre alors que la croissance dans les premiers mois de la vie est beaucoup plus générale.

Mais c'est surtout si on utilise le modèle de régression pour estimer la taille de quelqu'un de plus de 17 ans que l'on risque de commettre les erreurs les plus grossières. En utilisant un tel modèle un sujet de 40 ans devrait mesurer environ 285 centimètres !

$$40 * 4,944773 + 88,330923 = 286.121843$$

Il est parfaitement connu de tous, sans que cela ne soit nécessairement explicité et scientifisé que la taille des êtres vivants arrête d'augmenter lors de la maturité. J'ai volontairement pris cet exemple caricatural à titre d'introduction aux problèmes de contraintes en statistiques.

Pour fixer l'idée de départ voici les représentations graphiques des diagrammes de dispersion de la taille avec l'âge avec la droite de régression linéaire, dans le premier diagramme on se limite aux âges observés et dans le second on considère un intervalle plus grand pour mettre en évidence l'absurdité de l'extrapolation.

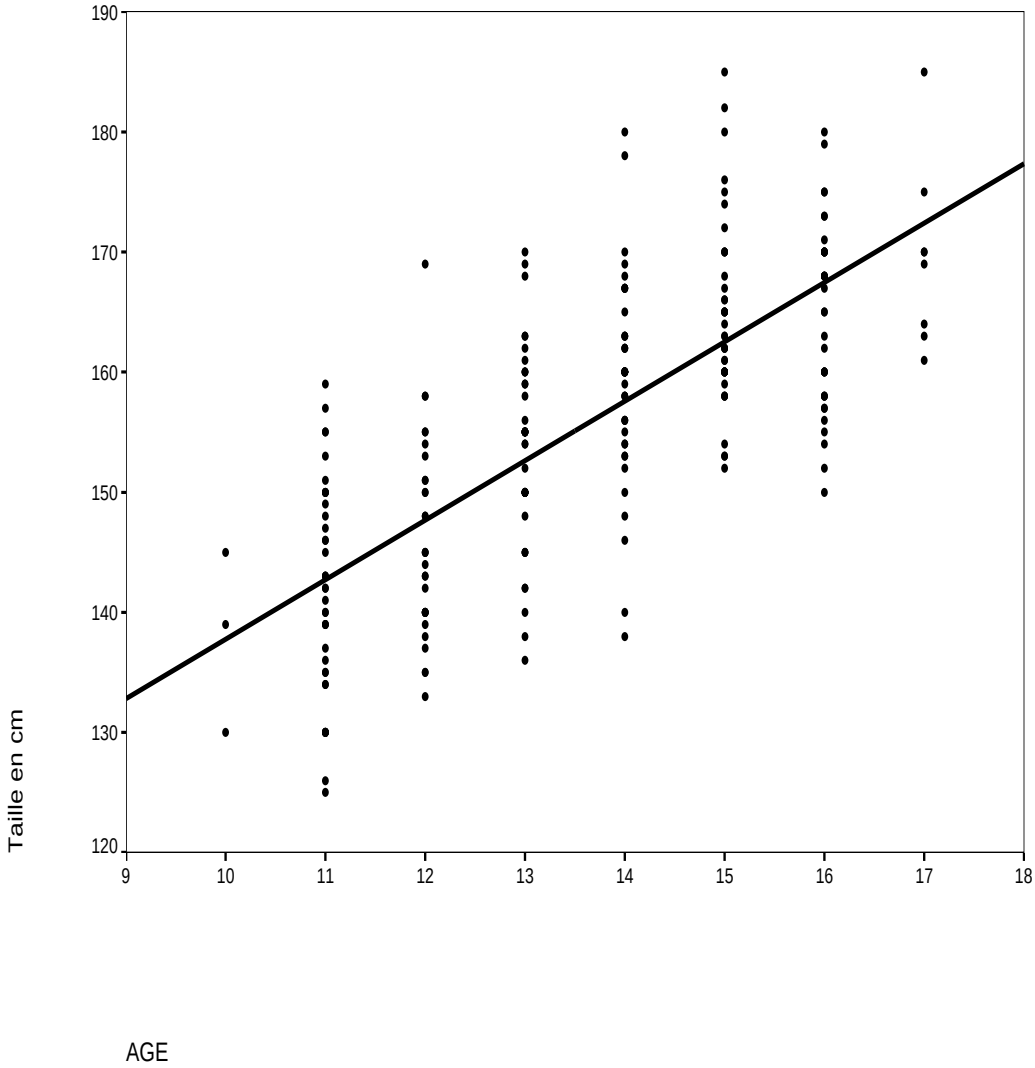

FIG. 1 – régression linéaire sans extrapolation

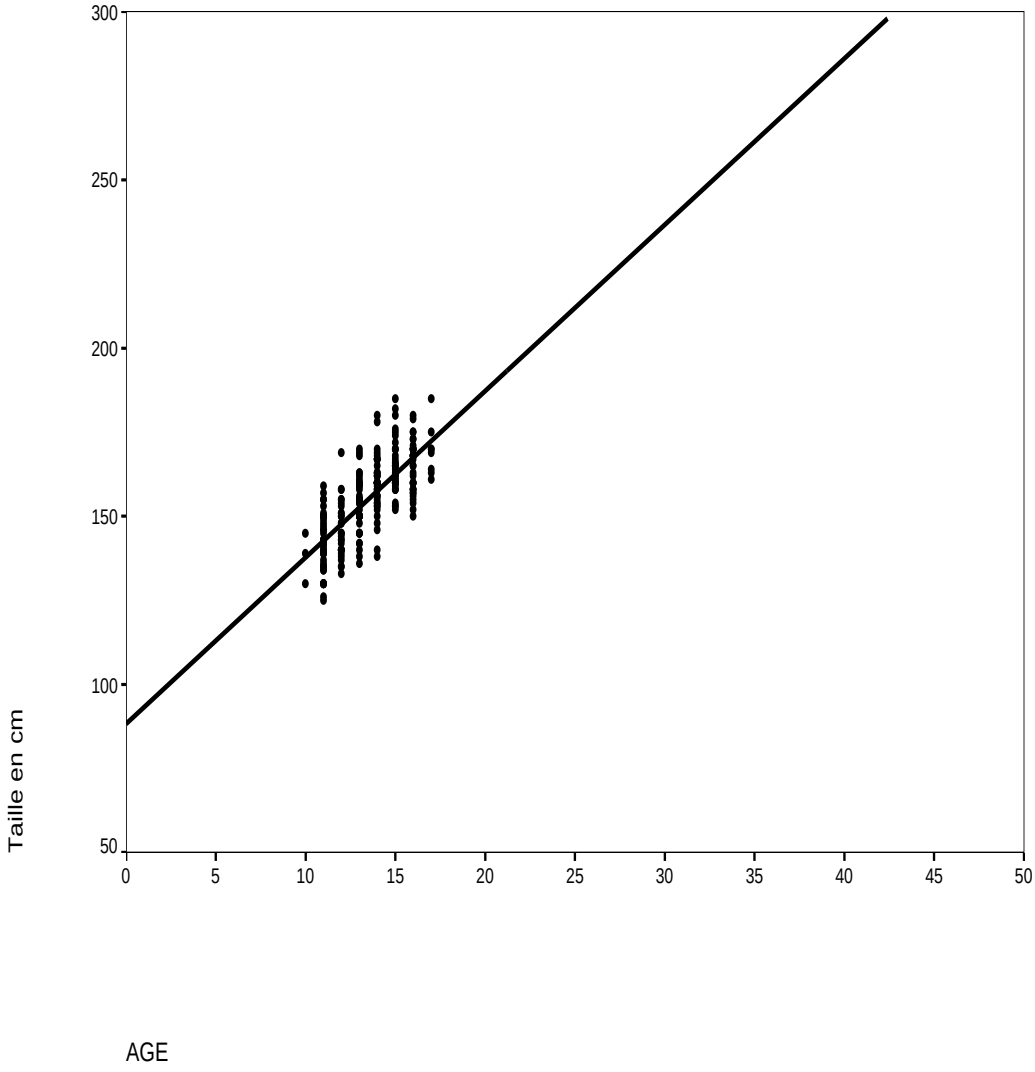

FIG. 2 – régression linéaire extrapolation aberrante

4 Nombre de licenciés, comptage

Dans le cas de la taille des adolescents nous avions une variable qui est mesurée en centimètres mais qui aurait pu être mesurée en millimètres voire en dixièmes de millimètre ou en micron, il s'agit d'une mesure d'intervalle qui correspond assez bien à ce que l'on appelle une variable continue. Le problème de limites à ne pas dépasser peut concerner des variables qui ne sont pas continues. Voici un message qui a été envoyé en Octobre 2001 par une lycéenne dans le forum de usenet fr.sci.maths . les messages de ce forum sont accessibles au monde entier avec des moteurs de recherche tel que google ou altavista. Ce forum de langue française est dédié aux sujets sur les mathématiques.

```
Subject: Problème de statistiques
Date: Fri, 19 Oct 2001 20:18:20

Voici un problème de statistiques de terminales S . Les statistiques
ne sont pas mon dada, pouvez-vous m'aider à le résoudre ?

Problème:

Années :                  1970      1980    1990    1996
Nombre de licenciés:       698      1554    1831    2056

On appelle Ti les années et Xi le nombre de licenciés en l'année Ti .

1. Représenter le nuage de points Ai ( Ti , Xi ) .
2. Calculer le coefficient de corrélation de ce nuage .
     Pouquoi cet ajustement affine est-il justifié ?
3. a) Ecrire l'équation de la droite de régression de X en T .
   b) Ecrire l'équation de la droite de régression de T en X .
4. a) Estimer le nombre de licenciés pour l'année 1998 .
   b) A partir de quelle année peut-on estimer que le nombre
      de licenciés dépassera 2,5 millions ?

La première question ne me pose pas de problème mais j'aimerais être
sure du résultat des autres questions.

                 Merci d'avance.
```

J'a envoyé une réponse à ce message et je vous livre cette réponse. Cette discussion a un rapport direct avec l'évolution de la taille des êtres humains, il y a une limite non dépassable

```
Subject: Re: Problème de statistiques
Date: 22 Oct 2001 12:20:10 +0200
Organization: Centre Interuniversitaire de Calcul de Toulouse


> 2. Calculer le coefficient de corrélation de ce nuage .

on doit trouver 0,966

>      Pouquoi cet ajustement affine est-il justifié ?


Il n'est pas vraiment justifié avec 4 observations


> 3. a) Ecrire l'équation de la droite de régression de X en T .

X= -98081,02 + 50,209 T


On peut déjà voir à partir de cette formule pourquoi l'ajustement
affine n'est pas vraiment justifié pour l'an 0 le nombre de licenciés
est négatif mais il le serait aussi en 1955. Pour des raisons
évidentes le nombre de licenciés ne doit pas être négatif et la
croissance ne doit pas être affine sur une longue période.


>     b) Ecrire l'équation de la droite de régression de T en X .


T=1955,514+ 0,01857 X
```

> 4. a) Estimer le nombre de licenciés pour l'année 1998 .

1998*50,209 -98081,02 = environ 2237

> b) A partir de quelle année peut-on estimer que le nombre de
> licenciés dépassera 2,5 millions ?

en appliquant la formule, on doit trouver l'an 48380, mais je trouve cela parfaitement absurde. Avec un tel modèle remonter avant 1955 est une bêtise faire une projection de plus de 100 ans dans le futur est à proscrire.

En tant que statisticien professionnel je me permets de critiquer un tel exercice de statistiques. Les statistiques devraient aider à mettre les mathématiques en contact avec une certaine réalité. C'est parce que les mathématiques appliquées sont enseignées à partir de tels exercices que certains pronostics sur l'évolution des cas de SIDA que j'ai vu en 1986 ou 1987 donnaient des millions de morts en France au début des années 2000 et plus de morts que d'habitants vers 2005.

À partir de telles données j'ai appris, à l'université, qu'il était légitime de faire de l'interpolation, c'est à dire regarder entre 1970 et 1996 mais pas de l'extrapolation c'est à dire regarder avant 1970 ou après 1996. Les marges d'erreurs sont énormes, mais on ne vous demande pas de les calculer...

Par ailleurs un nombre de licenciés, même si vous ne précisez pas, ne doit pas suivre un modèle affine par rapport au temps, mais plutôt une croissance rapide au début puis une asymptote, le nombre de licenciés ne peut dépasser la taille de la population de référence et celle ci ne peut pas devenir infinie. Einstein disait «Il n'y a que deux choses qui sont infinies l'univers et la bêtise humaine mais pour l'univers j'ai quelques doutes...»

Joseph Saint Pierre

Cette discussion concerne un problème qui, d'après moi, est inventé et dans lequel on a des limites à ne pas dépasser. Dans ce cas là le nombre de li-

cenciés varie unité par unité, mais le nombre de licenciés est nécessairement limité, pour des raisons différentes du cas de la taille mais pour des considération extérieures au problème posé.

Dans le cas du nombre de licenciés comme celui de la taille des êtres humains il y un problèmes de contraintes ; certaines mesures ne peuvent dépasser des valeurs données ou ne peuvent descendre en dessous de certains seuils. Beaucoup de méthodes statistiques simples comme la corrélation, ou la régression sont mises en oeuvre sans se soucier des problèmes de limites évoqués, mais cela est vrai pour la plupart des méthodes qui sont construites à partir de la corrélation telles que l'analyse en composantes principales ou l'analyse factorielle, parfois nommée analyse en facteurs communs.

5 Contraintes d'inégalités et Éventail du vivant

Je me suis rendu compte de l'importance de ce type de problèmes lorsque j'étais étudiant de maîtrise en 1978-1979 et lorsque j'ai été convié à faire un DEA, j'ai choisi d'étudier les contraintes d'inégalités en statistiques et j'ai fait une thèse dans laquelle les limites à ne pas dépasser avaient un rôle essentiel ; le titre de celle ci était : *Tests d'inégalité dans le modèle linéaire généralisé (Toulouse, 1983)*. Depuis je n'ai jamais cessé d'accorder une certaine importance à la gestion des limites à ne pas dépasser dans la modélisation statistique mais je ne me suis jamais livré, avant cet exposé, à une tentative de vulgarisation des idées de cette thèse. L'idée de faire un exposé de vulgarisation autour des problèmes de limites en statistiques m'est venu à partir de la lecture des livres de Stephen Jay Gould. Depuis de longues années je conseille la lecture de *La Mal-Mesure de l'Homme* de cet auteur pour tous les spécialistes de sciences humaines voulant faire des analyses factorielles, je considère ce livre comme très important ; en 1997 est paru, en français, une nouvelle version augmentée de *La Mal-Mesure de l'Homme* ainsi que *L'Eventail du vivant* un autre livre de Stephen Jay Gould très orienté vers les statistiques. Ce livre est sous-titré *Le mythe du progrès* et il est consacré aux problèmes de limites avec un intérêt particulier pour les sciences du vivant. Stephen Jay Gould n'est pas un statisticien, il enseigne la géologie, la biologie et l'histoire des sciences à l'université Harvard, c'est un excellent vulgarisateur, non seulement de ses disciplines de prédilection mais aussi des statistiques, à mon avis. Bien évidemment je ne peux que conseiller la lecture de *L'Éventail du vivant*. Ce livre existe, depuis, dans une collection bon marché, Points Sciences, cet exposé peut être, en partie considéré, comme

une incitation à la lecture de ce livre.

6 Limites dans le sport

Stephen Jay Gould est un passionné de sports, ce que je suis nullement, et il est particulièrement féru dans le base-ball, sport typiquement américain. Une partie importante de *L'Éventail du vivant* est consacrée au sport en général et au base-ball en particulier. Comme le souligne Stephen Jay Gould le base-ball est intéressant d'un point de vue statistique car les règles du jeu sont figées depuis longtemps, la performance des joueurs peut être mesurée individuellement et les archives des performances passées sont archivées par une société savante la Society for American Baseball Research cette société a donné naissance au mot «sabermetric» pour parler des statistiques faites sur des archives sportives en général.

Afin de donner une idée du style d'écriture de Stephen Jay Gould, même si il s'agit d'une traduction, je cite un court extrait de ce livre qui a l'avantage de concerner des sports plus répandus que le base-ball, l'avantage de présenter d'autres types de situations que purement sportives où l'extrapolation peut conduire à des conclusions farfelues. Par ailleurs il aborde un sujet qui concerne des thèmes de recherche que j'ai rencontrés souvent lors de traitements statistiques pour des universitaires, la comparaison entre les femmes et les hommes.

«Nombre de commentateurs ont observé que, pour une discipline donnée, la plupart des records féminins chutent plus rapidement que ceux des hommes, et que loin de se stabiliser, ils progressent de façon linéaire. Fait intéressant (voir Whipp et Wards 1992), les records dans la plupart des épreuves de course à pieds (du 200 mètres au 10000 mètres) ont progressé dans des proportions analogues - entre 5,69 et 7,53 mètres par minute et par décennie -, indépendamment de la distance. (Cette progression a été plus marquée pour le marathon avec 9,18 mètres par minute et par décennie. Cela conforte mon affirmation : cette épreuve est encore «immature» et l'évolution des records est encore linéaire - loin du mur de droite.) Si l'on regarde les résultats des femmes, les progrès sont compris entre 14,04 et 17,86 mètres par minute et par décennie (avec un bond fabuleux de 37,75 mètres pour le marathon).

Ces découvertes ont conduit à toutes sortes de spéculations, dont certaines sont passablement farfelues. Par exemple, en extrapolant leurs courbes , Whipp et Ward (1992) soutiennent que les femmes dépasseront les hommes dans pratiquement toutes les disciplines, et plus tôt qu'on ne le pense pour certaines d'entre elles. (Les courbes du marathon devant se croiser en 1998,

les femmes devraient battre les hommes sur cette distance).

Mais l'extrapolation est un jeu dangereux, généralement illégitime et souvent irréaliste. Je l'ai dit plus haut, en prolongeant les courbes suffisamment loin, on serait amené à conclure que toutes les distances seront un jour courues en un temps nul, voire négatif. (Ce même type d'extrapolation abusive donnent des chiffres aberrants souvent avancés pour la démographie planétaire : par exemple, dans quelques siècles, les êtres humains formeront une masse compacte égale au volume de la terre ; cette masse n'aura aucune possibilité de se répandre dans l'espace cosmique, car elle atteindra un taux de croissance si élevé que son diamètre augmentera à une vitesse supérieure à celle de la lumière ; or comme nous l'a appris Einstein, rien ne peut se déplacer plus vite que la lumière.) On atteindra les limites, les murs de droite et les progrès s'amenuiseront pour finalement s'arrêter.

Les femmes parviennent à surpasser les hommes dans certaines disciplines, par exemple, en natation sur de très longues distances, où la flottabilité et la répartition des graisses favorisent le physique et l'endurance des femmes (elles détiennent déjà le record absolu de la traversée à la nage de la Manche et de l'épreuve des Catalina - une quarantaine kilomètres entre Los Angeles et les îles Catalina -). Il pourrait en être de même pour le marathon mais je doute qu'un jour une femme puisse détenir un jour le record du monde sur 100 mètres plat ou celui d'haltérophilie en catégorie poids lourds. (Nombre de femmes battront toujours la plupart des hommes dans une discipline donnée - la plupart des femmes me battent dans pratiquement tout ce qui est physique. Mais n'oubliez pas que nous parlons ici de records mondiaux, mettant aux prises les plus grands champions ; et les particularités biomécaniques jouent ici un rôle déterminant.)

L'origine fondamentale des progrès plus rapides (et de la pente abrupte des courbes de records) accomplis par les femmes dans les disciplines sportives est évidente. Le principal coupable est le sexisme, sanctionné aujourd'hui par un heureux renversement de situation. Dans la plupart de ces disciplines, l'ouverture aux femmes est très récente. L'entrée des femmes dans l'univers du professionnalisme, avec son entraînement intense et sa compétition acharnée, date seulement de ces dernières années. Il n'y a pas si longtemps, les femmes étaient socialement conditionnées pour considérer la performance athlétique comme indigne de leur sexe. (une telle attitude est loin d'avoir totalement disparu) et nombre de grandes championnes du passé, durent affronter des railleries à l'encontre de leur excessive virilité. En d'autres termes, la plupart des courbes de records féminins en sont à leurs débuts - dans la phase linéaire,

à la pente très accentuée. Elles connaîtront un palier lorsque les femmes atteindront leurs propres murs de droite : et c'est seulement alors qu'il y aura une réelle égalité des chances. Pour l'instant, ces courbes témoignent de nos injustices passées et présentes.»

L'expression «mur de droite», très fréquemment employée par Stephen Jay Gould, désigne une limite maximale non dépassable. En ce qui concerne les limites des performances sportives la situation en 2002 a évolué depuis la parution en anglais du livre en anglais en 1996, le record mondial du marathon masculin a été battu en 1998, en 1999 et le 14 Avril 2002, soit trois fois, le record mondial du marathon féminin a été battu en 1998, en 1999, en septembre 2001 et enfin le 7 Octobre 2001 soit quatre fois. Depuis 1996, le record masculin s'est amélioré de 1 minute et 12 secondes, le record féminin s'est amélioré de 2 minutes et 19 secondes. Pour donner une perception de l'évolution du marathon, j'ai pris tous les records depuis que ceux ci sont homologués mais je n'ai gardé que le dernier record pour les années où le record a été battu plus d'une fois, ensuite j'ai fait une représentation graphique, un diagramme de dispersion, pour montrer l'évolution j'ai utilisé la méthode de lissage par les splines, disponible de façon évidente dans SPSS. Le record est mesuré en secondes, mais le but de cette représentation n'est pas de donner une valeur précise des temps du marathon mais l'allure de la courbe d'évolution du record pur les deux sexes. Il faut noter que le record mondial masculin du marathon n'a pas bougé entre en 1988 et 1998 soit pendant dix ans, et que le record mondial féminin du marathon n'a pas bougé entre 1985 et 1998 soit treize ans. Les raisons d'une telle stagnation puis d'une progression assez rapide sont difficiles à trouver, certains mauvais esprits pourraient voir la mise au point de nouvelles techniques de dopages. On peut imaginer aussi que certains sports qui ne nécessitent pas beaucoup de moyens matériels comme le marathon sont maintenant connus et pratiqués par des populations pour qui ce ne n'était pas le cas naguère, l'égalité d'accès aux sports entre les femmes et les hommes n'est pas la même suivant les pays. La détentrice du record actuel est une Kenyane, Catherine Ndereba la première femme issue d'un pays du tiers monde à détenir le record du marathon, l'a obtenu en 1998 il s'agissait de Tegla Loroupe, une autre Kenyane. Pour les hommes, il faut remonter à 1960 avec l'Éthiopien Abebe Bikila aux jeux olympiques de Rome, il avait couru pieds nus, il était le premier Africain à gagner une médaille olympique, et il est arrivé sous l'arc de Triomphe d'où, 25 ans plus tôt, les troupes de Mussolini étaient parties pour envahir l'Éthiopie. Tous ces faits avaient rendu cette victoire hautement symbolique, comme l'écrit Stephen Jay Gould l'histoire des performances sportives témoignent des injustices, c'est vrai pour l'injustice entre les sexes et c'est vrai aussi pour

l'injustice entre les continents et ces deux types d'injustice sont entremêlées.

Le problème des limites dans le sport a été largement médiatisé en raison de très nombreux cas de dopages, mais aussi en raison des sommes d'argent qui sont en jeu dans le sport et tout ce qui tourne autour. Il existe des critiques contre une évolution qui mêle la recherche des performances et une certaine aliénation des sportifs et des supporteurs. Après avoir commenté un livre qui est fortement consacré au sport sans aborder ce type de critique, je conseille la lecture du livre *Le sport contre les peuples*, du professeur de philosophie Toulousain Robert Redeker.

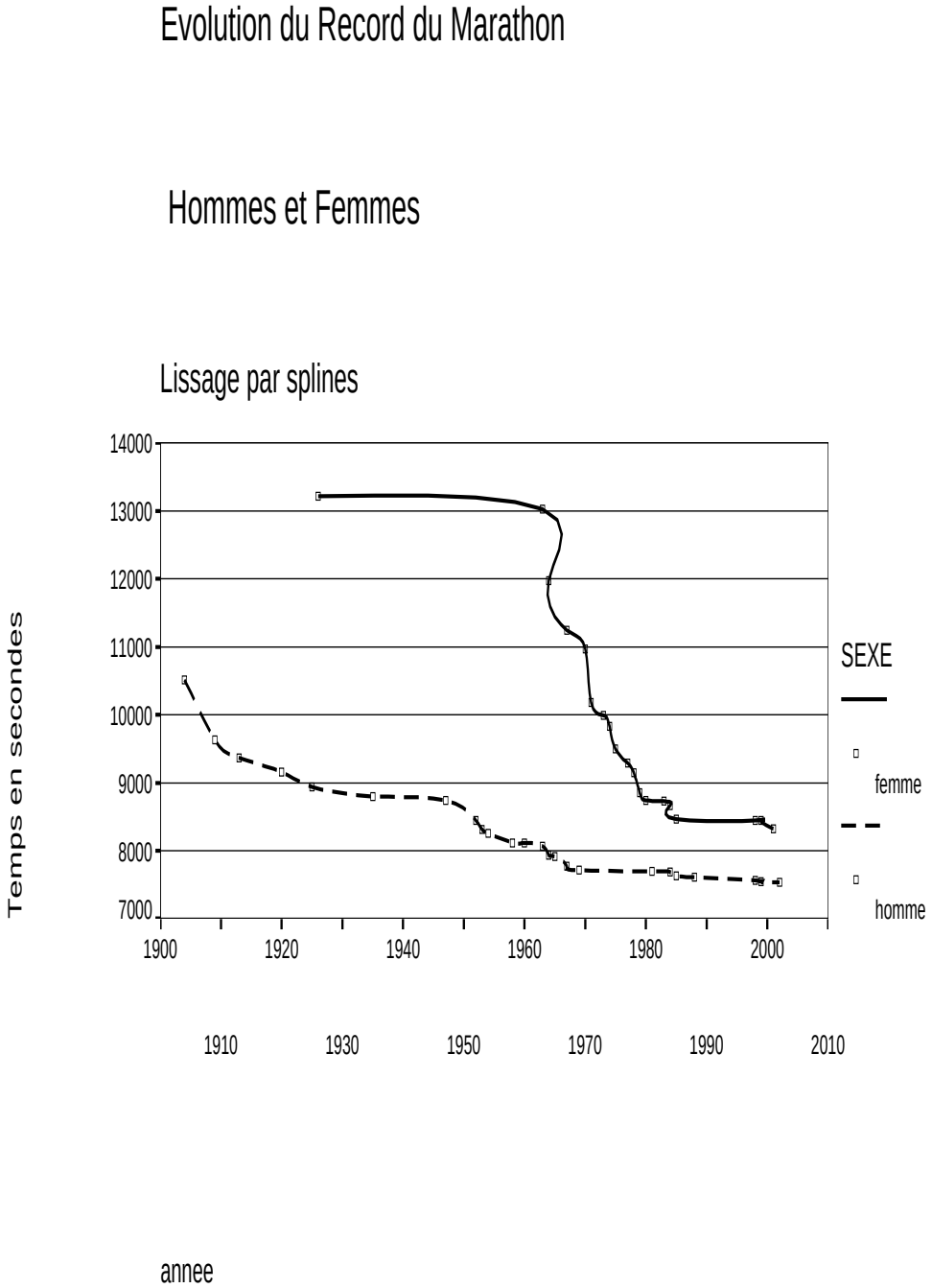

FIG. 3 – évolution du record du marathon dans le temps

La courbe obtenue par les «splines» permet de donner une idée de l'évolution du record à venir pour le marathon. Les calculs qui sont effectués pour obtenir cette représentation sont relativement compliqués et il est hors de question de donner une explication détaillée. Il s'agit en l'occurrence d'une méthode utilisée de manière purement descriptive et donc exploratoire. La méthode permet de percevoir une limite à la progression du record du Marathon mais ne permet pas de fixer la valeur de cette limite ni même la date où cette limite pourrait être atteinte.

Afin de montrer, le défaut de l'extrapolation par un modèle de régression linéaire je me suis limité aux observations entre 1960 et 1985 et j'ai effectué une régression linéaire sur les hommes et sur les femmes. Ces deux régressions montrent qu'en extrapolant les tendances linéaires observées sur cette période de 25 ans pour les années entre 1985 et 1990, cela aurait prédit de meilleures performances pour les femmes à partir de 1988 alors que ce n'est pas encore le cas.

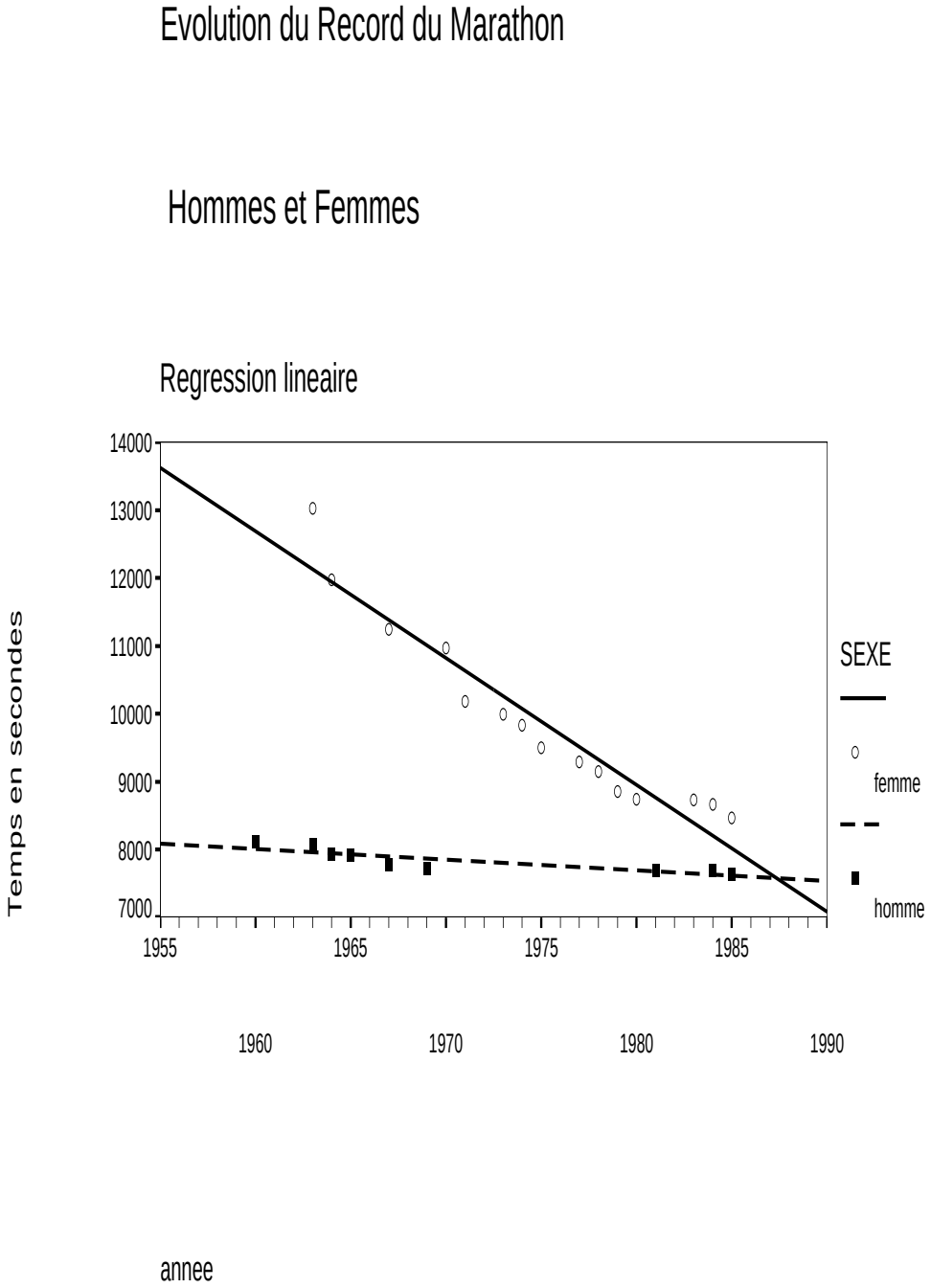

FIG. 4 – évolution du record du marathon dans le temps

7 Limites infranchissables en physique

Le problème des limites infranchissables est connu dans le domaine de la physique depuis assez longtemps. le premier exemple concerne la température et date du dix-neuvième siècle. En 1851 les travaux de thermodynamique de William Thomson (1824-1907) ont mis en évidence qu'il y avait un zéro absolu, cela valut l'anoblissement à William Thomson par la reine Victoria sous le nom de Lord Kelvin. L'existence d'un zéro absolu a amené la création d'une nouvelle unité de mesure de la température nommée le Kelvin en l'honneur de Lord Kelvin cette unité est celle du système international et se note K. Le zéro absolu correspond à 0 K et à $-273,15°$ Celsius c'est à dire l'échelle ordinaire des thermomètres usuels et de la météorologie française par exemple. Pour passer de la température en Celsius à la température en Kelvin il suffit d'ajouter $273,15$, l'eau gèle à $273,15$ K et bout à $373,15$ K. La conversion est plus facile que le passage entre degrés Farenheit et degrés Celsius. Il est impossible d'avoir une température en dessous de 0 K mais il est aussi impossible d'atteindre cette température.

L'autre exemple concerne les vitesses, les travaux d'Einstein sur la relativité au début du vingtième siècle ont montré que la vitesse de la lumière, soit environ 300000 kilomètres par seconde, était infranchissable et même inatteignable par les objets ordinaires

Les deux cas de la température et de la vitesse de la lumière sont abordés par Jean-Marc Lévy-Leblond dans le livre «La pierre de touche», sous titré «La science à l'épreuve...», l'auteur donne une appréciation très intéressante sur la nature apparemment finie des limites de $-273,5°$C et $300000km/s$. D'après l'auteur la nature finie des limites est liée à des conventions de mesures qui sont historiquement justifiables mais qui n'empêche pas les valeurs limites d'être conceptuellement infinies.

Le cas de la température est peut être plus facile à percevoir que celui de la vitesse car les imperfections de l'échelle linéaire de température est perceptible de façon plus immédiate. Certains thermomètres fonctionnent avec du mercure, ce métal, liquide aux températures usuelles, se dilate sous l'effet de la chaleur l'importance de la dilatation est proportionnelle à l'augmentation de la température, c'est à dire qu'il y a une liaison linéaire entre l'augmentation de la température et l'augmentation du volume du mercure dans le thermomètre et donc avec l'augmentation de la longueur de mercure dans le tube gradué où on peut lire la température. Cette liaison linéaire est presque parfaite pour les thermomètres médicaux à mercure étalonnés entre

35°C et 42°C pour plus de facilité on peut même faire en sorte qu'une hausse de 1°C corresponde à 1 centimètre d'allongement de la barre de mercure ; donc 1 millimètre correspond à $0,1$°C. Ce principe d'élongation régulière d'un liquide sous l'effet de la chaleur n'est valable que dans un certain intervalle, lequel dépend du liquide considéré. Par exemple il serait difficile de faire un thermomètre avec de l'eau, l'eau a sa densité maximale à 4°C, cela veut dire que si la température monte au dessus de 4°C l'eau se dilate, mais si la température descend au dessous de 4°C elle se dilate aussi, cela est heureux pour les poissons qui peuvent vivre dans l'eau sous les couches de glace qui se concentrent à la surface, cela oblige à prendre un liquide dont la température de fusion et d'ébullition sont éloignées des températures que l'on veut mesurer ; le mercure gèle vers -30°C et bout vers 356°C ; cela veut dire que l'usage d'un thermomètre à mercure serait impossible pour des froids sibériens et même peu adapté à contrôler la température d'un congélateur familial... Il n'y a pas de thermomètre universel et les problèmes de la mesure des températures montrent bien la nature conventionnelle des mesures employées.

8 Régression sur les pourcentages

Parmi les problèmes de limites impossibles à dépasser et très communes on peut citer les pourcentages dans certaines situations. Le cas présenté un peu plus haut concernant le nombre de licenciés aurait pu être présenté en terme de pourcentages par rapport à une population donnée, le nombre de licenciés ne peut descendre en dessous de 0% et ne peut pas dépasser 100% de la population de référence, il y a dans ce cas des limites inférieures et supérieures dans la variation possible, dont la justification est intuitivement plus facile à comprendre que le zéro absolu comme limite inférieure de la température ou que la vitesse de la lumière comme limite supérieure de la vitesse.

Comme exemple de situation récente où une approche linéaire à conduit à des erreurs de prévisions assez remarquables je peux citer la fabrication de téléphones portables. En l'espace de quelques années les habitants des pays riches se sont massivement dotés de téléphones portables et le marché a connu une croissance phénoménale, les chaînes de fabrication ont fonctionné à plein régime. La croissance de ventes et de fabrication ne pouvait pas être linéaire, malgré les pertes, les vols, les détériorations, les collectionneurs le marché ne pouvait que se saturer et le nombre de téléphones vendus

ne pouvait que se stabiliser à un niveau plus bas que celui correspondant à une période du passage de 0% de la population possédant un portable à plus de 60%. La croissance prévue par certains constructeurs était exponentielle...

En période d'élections les résultats électoraux, mais aussi les sondages, fournissent énormément de pourcentages aux media ; il est fréquent d'entendre des raisonnements utilisant implicitement des extrapolations linéaires sur des pourcentages. À titre d'exemple je présente une régression linéaire fait sur les pourcentages de voix obtenus par le Parti Communiste Français lors des principales consultations électorales Françaises entre 1980 et le premier tour de l'élection présidentielle de 2002. On voit facilement à partir de la droite de régression qu'à partir de 2015 environ le pourcentage de voix obtenues par le PCF devrait être inférieur à 0% ce qui est, bien évidemment, absurde.

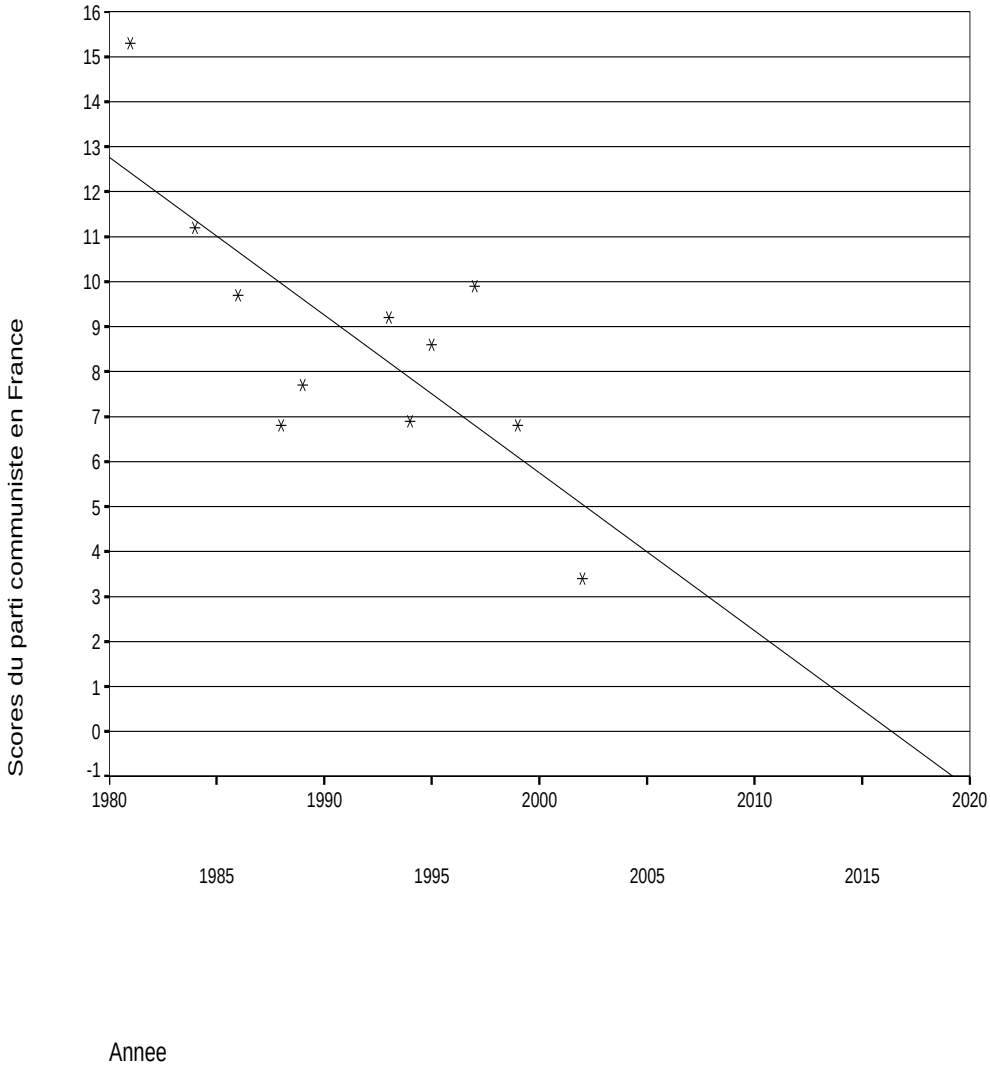

FIG. 5 – régression linéaire sans extrapolation

9 La régression logistique

Il est possible de démultiplier à loisir les exemples de situations où la régression faite sur les pourcentages conduit à des extrapolations absurdes, où les limites normalement infranchissables de 0% et 100% sont dépassées. Il est possible faire une analogie avec la remarque de Jean-Marc Lévy-Leblond concernant la température du zéro absolu et la vitesse de lumière et précisant qu'il s'agit de limites conceptuellement infinies. On peut dire que dans bien des situations 0% et 100% sont valeurs conceptuellement infinies et il est possible de transformer par de nombreuses fonctions mathématiques une échelle allant de 0 à 100 en une échelle allant de $-\infty$ à $+\infty$ de façon schématique c'est l'idée qui est utilisée dans la régression logistique la fonction qui est utilisée est $log(x/(1-x))$. Il faut noter que d'un point de vue mathématique 1 et 100% désignent la même chose. La régression logistique peut être vue dans le cas particulier des pourcentages un moyen mathématique de systématiquement éviter le problème présenté dans cet exposé c'est à dire le dépassement de limites normalement infranchissables. Cet exposé n'a pas pour but de présenter la théorie de la régression logistique ni même d'expliqer sa mise en oeuvre avec un logiciel mais plutôt de montrer une de ses justifications et de donner l'idée de l'artifice mathématique utilisé qui peut être éventuellement adapté à d'autres situations. Si une limite non franchissable est connue à priori il est légitime de la considérer comme conceptuellement infinie et de chercher une fonction mathématique qui renvoie cette valeur à l'infini, bien évidemment le choix de la fonction mathématique se fait souvent en raisons de facilités de calculs mais aussi d'autres justifications difficiles à expliciter.

10 Conclusion

Il n'est pas possible de donner une méthodologie générale pour résoudre les problèmes de limites à ne pas dépasser en statistiques. Le cas de la régression logistique est un peu particulier. Il sûr qu'il faut toujours faire appel au bon sens et à la prudence avant d'avancer un résultat physiquement absurde qui découlerait d'extrapolations mathématiques aventureuses. Il est toujours possible qu'un modèle statistique représente une vérité locale correspondant au domaine, à la période, à la population observé... On peut songer à utiliser des méthodes statistiques plus compliquées mais il est nécessaire pour cela de comprendre la nature du problème de départ, cela peut être résumé par la citation de Bachelard : «Il faut comprendre pour mesurer et non mesurer pour comprendre». La solution la plus simple consiste à ra-

mener une extrapolation dépassant une limite infranchissable à cette valeur limite mais ce n'est pas toujours satisfaisant ni réaliste, l'exemple du 0 absolu dans la température le montre. Bien évidemment on ne peut que conseiller la prudence dans l'usage des statistiques et des nombres en général, surtout lorsque les nombres se détachent trop nettement des réalités qu'ils sont censés représenter.

Ce document a été écrit rapidement à partir de notes pour un exposé dans le cadre des Écoles Doctorales de l'Université de Toulouse le Mirail. je compte sur les questions et les commentaires que je recevrais pour éventuellement en faire un document un peu plus achevé que je mettrais sur ma page web.

Table des matières

1 Mise en garde **2**

2 Préambule **2**

3 Taille d'adolescents, un cas continu **4**

4 Nombre de licenciés, comptage **9**

5 Contraintes d'inégalités et Éventail du vivant **12**

6 Limites dans le sport **13**

7 Limites infranchissables en physique **20**

8 Régression sur les pourcentages **21**

9 La régression logistique **24**

10 Conclusion **24**