

S. KOUADI

Analyse d'un facteur de risque : tabac et cancer bronchique

Les cahiers de l'analyse des données, tome 9, n° 4 (1984),
p. 447-456

http://www.numdam.org/item?id=CAD_1984__9_4_447_0

© Les cahiers de l'analyse des données, Dunod, 1984, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

ANALYSE D'UN FACTEUR DE RISQUE : TABAC ET CANCER BRONCHIQUE

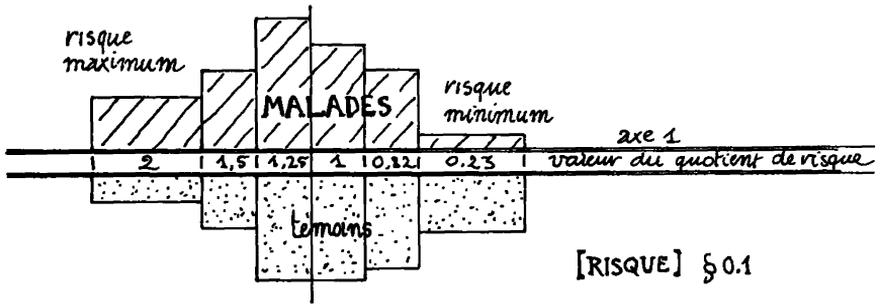
[RISQUE]

par S. Kouadi (Madame Madhy)

0 La définition des facteurs de risque : Tous les sujets ne sont pas également susceptibles de contracter une maladie donnée ; et s'ils la contractent ils y résistent inégalement. Outre la contagion et les dispositions naturelles, interviennent des habitudes professionnelles, alimentaires et autres. Ainsi il est admis et prouvé que l'usage du tabac intervient dans la genèse des cancers bronchiques. Avant de présenter quelques résultats des analyses déjà effectuées sur les données d'une enquête consacrée à l'usage du tabac chez quelque 3000 sujets dont 1/3 de cancéreux (cf. § 1), il semble utile de suggérer comment l'analyse statistique de données épidémiologiques peut définir des *facteurs de risque*.

0.1 Une maladie : Partons d'un cas simple. On considère l'incidence d'une maladie déterminée sur une population présumée homogène. On suppose que le risque de m est lié à un ensemble \mathcal{O} de variables. Selon l'usage nous notons J_q l'ensemble des modalités d'une variable q ; et J (réunion des J_q) l'ensemble des modalités de toutes les variables. A un ensemble I d'individus est associé un tableau de description $I \times J$ sous forme disjonctive complète (en $0, 1$: $k(i, j) = 1$ si le sujet i a la modalité j ; et zéro sinon). On suppose que I comprend un sous-ensemble I_m de malades et un sous-ensemble I_t de témoins (i.e. de sujets exempts de la maladie m). Si comme on peut l'espérer..., la fréquence de la maladie m n'est pas très élevée dans la population étudiée, un échantillonnage uniforme au sein de celle-ci conduirait à un sous-ensemble I_m d'effectif faible relativement à celui de I_t : disproportion qui ne paraît être aucunement être favorable à la comparaison projetée. Il semble donc préférable de dissocier deux problèmes statistiques : le premier relativement simple, est l'estimation approchée de la fréquence de m au sein de la population visée ; le second, plus délicat et objet propre de notre étude, est de déterminer comment cette probabilité varie en fonction des modalités de J . Pour ceci le plus simple nous semble être de disposer d'ensembles I_m et I_t d'égal effectif - ou, ce qui revient au même de donner aux individus des deux ensembles des pondérations relatives telles que I_m et I_t aient même poids total. Le cumul des sous-ensembles de lignes I_m et I_t du tableau $I \times J$ donne un tableau à deux lignes, qu'on notera i_m et i_t : $k(i_m, j)$ étant le nombre de malades rentrant dans la modalité j ; et de même $k(i_t, j)$ le nombre de témoins rentrant dans cette modalité. L'analyse factorielle de ce tableau à deux lignes fournit un axe unique, sur lequel il est possible de projeter en éléments supplémentaires les deux sous-ensembles I_m et I_t .

(1) Docteur 3° cycle ; le présent travail expose les résultats d'une thèse soutenue le 6 avril 1981 devant un jury de l'Université Pierre et Marie Curie.



En découpant l'axe en une suite d'intervalles convenablement choisis, on construit les deux histogrammes des sujets et des témoins : le rapport de la hauteur des malades, à celui du créneau des témoins, afférents à un même intervalle, donne pour cet intervalle un coefficient de correction à la probabilité de la maladie m , mesurée préalablement sur l'ensemble de la population visée. Ainsi le facteur (au sens précis qu'a ce terme en a. factorielle) issu de l'analyse du tableau à deux lignes $\{im, it\} \times J$, s'interprète comme un facteur de risque (au sens épidémiologique du terme) ; le risque étant évidemment le plus grand côté où se projette le point im .

La construction effectuée ici, n'est autre que celle de la discrimination par affectation barycentrique, utilisée dans [AORTE] (cf. Nakache et coll ; in Cahiers Vol VII n° 4 p. 427, 1977) et exposée dans [SEP. CORR.] (même cahier p. 381). La discrimination, et par le fait même le calcul du risque, pouvant d'ailleurs être rendus précis si (à condition que les effectifs de l'échantillon le permettent) on effectue pour chaque intervalle de l'axe (ou du moins ceux assez chargés de points) une analyse factorielle particulière restreinte aux seuls individus qui s'y projettent (plus exactement que l'on calcule d'après ceux-ci un tableau à deux lignes im et it , sur lequel on opère comme précédemment.

0.2 Plusieurs maladies : Une investigation ne portera pas généralement sur une maladie m unique mais plutôt sur un domaine étroit de la pathologie, subdivisé en un ensemble M de maladies entre lesquelles le diagnostic est plus ou moins facile. On considèrera alors un ensemble I d'individus, comprenant un sous-ensemble It de témoins et des sous-ensembles Im afférents à chaque maladie. Quant aux pondérations, il semble naturel de donner même poids aux témoins qu'à la totalité des malades ; la part assignée à chaque maladie dépendant de l'attention accordée à chacune de celles-ci et aussi des données disponibles (une maladie dont on n'a que quelques cas ne pouvant figurer ou en éléments supplémentaires ; tandis qu'aux maladies bien représentées on peut attribuer des poids égaux... ou proportionnels à leurs probabilités respectives. Le tableau $I \times J$ des cas individuels étant construit, on lui associe comme au § 0.1, un tableau $ID \times J$ où $ID = \{im | m \in M\} \cup \{it\}$, est l'ensemble des diagnostics possibles (l'une des maladies m de M ; ou témoin t) ; une ligne im étant calculée par cumul des lignes des malades atteints de m (et de même pour it). Ici encore, on analysera le tableau $ID \times J$: mais ce tableau ayant plus de deux lignes fournira plus d'un facteur. On peut toutefois présumer que le premier facteur sera un facteur général de risque ; tandis que le deuxième facteur distinguera les diagnostics. Des histogrammes, construits pour des sous-populations éventuellement définies à la fois par des diagnostics et des régions du plan 1×2 (e. g.

malades ayant un facteur F1 négatif...), permettant de calculer des coefficients de risque proprement dit.

0.3 L'âge : Parmi les facteurs de risque, il en est un auquel finalement nous succombons tous : l'âge. En un certain sens, ce facteur est même omniprésent, sous-tendant les autres. Dans beaucoup de cas on ne peut éviter que les variables retenues comme pertinentes ne soient normalement corrélées à l'âge. Ainsi face à diverses affections, l'hypertension est un facteur de risque reconnu ; mais d'autre part la tension artérielle s'accroît communément avec l'âge : en sorte qu'une même tension inquiétante chez un jeune homme n'est pas notable chez un homme d'âge mûr. Si l'on n'y prend garde des analyses telles que celles suggérées ci-dessus (§§ 0.1, 0.2), peuvent être dominées par l'âge et n'aboutir qu'à une seule conclusion sûre, claire, mais peu intéressante : le risque augmente avec l'âge.

Une première défense radicale, est de borner l'analyse à des individus rentrant tous dans une tranche d'âge assez étroite : e.g. de 40 à 50 ans. Mais il est également possible de conserver l'ensemble des individus de tous âges dans une même analyse. Nous rappellerons ici comment procèdent F. Benjelloun et L. Guize (cf. [ANA. MED.] § 4.1 ; *in* Cahiers Vol V n° 3, 1980). Deux voies sont explorées : d'une part tenir compte de l'âge dans le codage : au lieu e.g. de trois modalités "tension basse-moyenne-haute" prendre "tension basse-moyenne-haute, relativement à la moyenne de sa classe d'âge" ; d'autre part interpréter non la position elle-même du sujet sur le diagramme (e.g. plan 1 x 2) issu de l'analyse factorielle, mais l'écart du sujet par rapport au centre de gravité de sa classe d'âge (cet écart étant dans le sens d'un accroissement ou d'une diminution du risque ; selon que le vieillissement est prévu ou différé).

Dans la présente étude, plusieurs des variables recueillies pour apprécier l'usage du tabac, dépendent de l'âge : un sujet âgé de moins de 40 ans ne peut guère avoir fumé 35 ans ou plus ; le nombre total de cigarettes fumées s'accroît au cours du temps. Certains usages - cigarettes roulées à la main, cigarettes à bout filtre - ne sont pas également distribuées dans toutes les générations. Cependant, pour délicate qu'en soit l'élaboration, les données méritaient d'être analysées ; et même si des analyses ultérieures plus précises semblent souhaitables, des résultats frappants ont déjà été obtenus.

1 Les données disponibles

1.1 Structure de l'enquête : A l'initiative du National Cancer Institute des Etats-Unis, une enquête internationale a été entreprise sur l'incidence de l'usage du tabac sur le cancer bronchique. Notre travail porte exclusivement sur des données recueillies en France dans plusieurs hôpitaux à Paris, Bordeaux, Marseille et Nancy.

Sur les 3026 sujets de l'enquête il y a 2887 hommes (95%) et 139 femmes (5%). Voici pour les hommes la composition en cancers et témoins, fumeurs et non fumeurs.

	Non-fumeurs	fumeurs
cancers	11	941
témoins	368	1536

Il faut préciser comment ces sujets ont été recrutés. Les cas de cancers sont pris au fur et à mesure de leur admission dans les services hospitaliers participant à l'enquête : le seul critère étant qu'il s'agisse d'un cancer bronchique d'un type ou d'un autre (cf. *infra*). Les témoins sont des malades hospitalisés pour des causes

sans rapport avec les bronches ni le tabac, de manière à ne pas biaiser l'étude ; et ils sont choisis dans la proportion approximative de deux témoins pour un cancer, de telle sorte que les profils d'âge (et aussi, dans une certaine mesure, d'activité professionnelle) des deux sous-échantillons, soient les mêmes). Il semblait *a priori* préférable de prendre des témoins dans la rue, pour qu'ils soient aussi parfaitement que possible représentatifs de leur classe d'âge (et de leur profession) : mais la passation du questionnaire (complétée d'examens divers) étant assez longue, on a dû se contenter de sujets disponibles parce qu'hospitalisés.

Le questionnaire complet couvre douze pages. Il est composé de quatre parties qu'on notera Q1, Q2, Q3, Q4.

La première partie Q1 est indépendante de l'état pathologique du malade : elle concerne l'état civil, l'âge, la situation professionnelle.

La partie Q2 concerne l'usage du tabac ; en distinguant entre cigarette, cigare, cigarillo, pipe, tabac à priser et à chiquer ; les modalités d'usage de la cigarette (inhalation de la fumée etc.) étant particulièrement détaillées.

La partie Q3 est relative aux maladies antérieures du sujet.

La dernière partie Q4 concerne la consommation d'alcools et autres boissons éventuellement excitantes ; avec des données d'examens de laboratoire (taux de cholestérol, groupe sanguin) ; et des indications sur le régime alimentaire (beurre, fromages, carottes, bananes...) et la prise de vitamines en comprimés.

A chaque questionnaire est joint le diagnostic ; distinguant entre les témoins, six types de cancers bronchiques définis d'après une nomenclature internationale (classification de Kryberg) :

Sigle	Libellé	taux
TEM	témoin	
KEPI	cancer épidoïde	73%
ANAP	carcinome anaplasique à petites cellules	7%
ANAG	carcinome anaplasique à grosses cellules	3%
ADEN	adénocarcinome	7%
INDI	indifférencié	10%
KMIX	cancer mixte	e (3 cas)

N.B. Dans ce tableau les taux sont ceux calculés relativement aux seuls malades.

Nous ne donnons pas pour l'instant le détail du questionnaire nous réservant d'en citer diverses questions ou modalités de réponse dans la suite de l'exposé.

1.2 Les tableaux analysés : Outre le diagnostic et l'âge (ainsi que quelques autres éléments de la partie Q1), nos analyses portent essentiellement sur la partie Q2, plus précisément sur les modalités d'usage de la cigarette ; mode de consommation du tabac prédominant et dont les calculs élémentaires de corrélation ont montré qu'il était le plus lié au cancer.

D'emblée une première constatation apparaît : le cancer du poumon atteint quasi exclusivement des fumeurs : d'après notre échantillon la fréquence de ces cancers est environ 20 fois plus importante chez les fumeurs que chez les non-fumeurs :

$$(941/1536)/(11/368) = 20,5$$

Même si un nombre tel que 11 est en butte à des fluctuations d'échantillonnage importantes, il s'impose de conclure que l'usage du tabac est une condition quasi nécessaire (non une condition suffisante...) de l'apparition d'un cancer bronchique. Cette conclusion est encore corroborée si l'on considère que les non-fumeurs atteints de cancer du poumon présentaient un terrain pulmonaire propice du fait de leurs antécédents pathologiques.

Ceci étant vu, il reste à préciser comment chez les fumeurs eux-mêmes, (plus précisément les fumeurs de cigarettes) le taux de cancer varie en fonction des modalités d'usage du tabac : durée, fréquence, inhalation. Nous disons variation de taux : les taux eux-mêmes sont en dehors du domaine de l'enquête qui n'a pas échantillonné l'ensemble de la population (éventuellement découpée en strates d'âge, profession etc...) mais seulement associé des témoins à des malades. C'est pourquoi ci-dessus nous avons seulement dit : risque pour les fumeurs multiplié par 20 relativement aux non-fumeurs.

D'autre part, il apparaît sur notre échantillon que le cancer bronchique est une affection frappant beaucoup plus souvent les hommes que les femmes : avec moins de 50 cas de cancers chez la femme, une analyse statistique semble même impossible. Toutes les analyses que nous avons faites, sauf la première portent donc exclusivement sur les hommes.

Il est maintenant facile de définir approximativement en peu de mots les tableaux que nous avons analysés.

Le premier T1 est un tableau $I \times J$ sous forme disjonctive complète croisant l'ensemble I de tous les sujets fumant la cigarette (hommes ou femmes ; malades ou témoins) avec un ensemble J de modalités qui mis à part l'âge, le sexe et le diagnostic concernent l'usage de la cigarette.

Les autres tableaux (de T2 à T6) sont des sous-tableaux de Burt croisant un ensemble ID de diagnostics avec divers ensembles de modalités relatifs à la cigarette et comportant ou non explicitement l'âge ; celui-ci intervenant de toute façon, par la quantité de tabac fumée (quantité estimée...) la durée d'usage ; etc. . Selon le cas, interviennent tous les diagnostics ou seulement les formes de maladies (à l'exclusion des témoins). Toujours (cf. § 0.2) le nombre $k(i,d,j)$ du tableau, est défini comme le nombre des sujets fumeurs masculins i rentrant à la fois dans le diagnostic d et la modalité j . A ces sous-tableaux de Burt, il est possible (cf. § 0,1, § 0.2) d'adjoindre en lignes supplémentaires les sujets individuels et d'en préciser par des histogrammes la répartition sur les axes.

Dans la suite nous donnons des extraits des résultats d'analyse. Un exposé plus complet se trouve dans la thèse.

2 Résultats d'analyse : Nous considérons successivement trois analyses. La première (§ 3.1) a été effectuée sur un tableau en $(0,1)$ (forme disjonctive complète) ; (les autres, sur des sous-tableaux de Burt comportant l'un sept (§ 3.2), l'autre cinq diagnostics (§ 3.3).

2.1 Analyse d'un tableau de description logique : Dans cette analyse dix questions ont fourni les modalités principales ; quelques modalités supplémentaires avaient été introduites, mais nous n'en parlerons pas puisqu'elles n'ont rien apporté de plus à l'interprétation. Voici le texte des questions, avec les modalités retenues.

Sexe : H , F ;

Nombre d'années pendant lesquelles le sujet a fumé : de AF1 (< 5) jusqu'à AF7 (>55).

Nombre de cigarettes fumées par jour: -10 (<10), -30 (de 10 à 30), +30 (i.e. >30).

Diagnostic : K (tous cancers), T (témoins).

Age au diagnostic : A < 49 ; A < 59 (i.e. de 50 à 59) ; A < 69 ; A > 70.

Inhalation : INAL (inhale) ; INAP (n'inhale pas).

Situation du fumeur : FACT (fume actuellement) ; EXFU (ex-fumeur)

Type de cigarettes : BRUN, BLON.

Usage de cigarettes avec ou sans filtre : FILT, SFLT (sans filtre), F-SF (les deux types).

Changements de marque : de MAR1 (sans changement) à MAR4 (3 changements).

Dans le plan des axes 1 x 2, on note d'abord l'opposition entre témoins et cancers qui se fait principalement dans la direction de l'axe 1. Du côté des témoins (axe 1 négatif) on trouve toute sorte de variables associées à l'usage modéré du tabac : e.g. EXFU (ex-fumeur) ; -10C (moins de 10 cigarettes par jour) ; AF1 a fumé moins de 5 années, INAP (n'inhale pas la fumée) ; les modalités opposées allant avec les cancers (K). Il est intéressant de suivre le long de l'axe 2 (de l'extrémité négative à l'extrémité positive) deux chaplets de modalités associées à l'âge : l'âge lors du diagnostic, de A > 70 (plus de 70 ans) à A < 49 (moins de 49 ans) ; et le nombre d'années pendant lesquelles le sujet a fumé : de AF7 (durant plus de 55 années) à AF3 (de 15 à 24 années). Quant aux modalités AF2 (de 5 à 14 années) et AF1 (moins de 5 années), elles se rabattent du côté négatif de l'axe 1 : en effet compte tenu de ce que les sujets de notre enquête (sujets atteints de cancers bronchiques, donc aussi sujets appariés avec eux) ont rarement un âge inférieur à 40 ans, les modalités AF1 et AF2 ne peuvent guère comprendre que des sujets ayant cessé de fumer depuis plusieurs années, (ou ayant commencé de fumer après 25 ans), c'est-à-dire des sujets ayant fait du tabac un usage plutôt modéré. On notera que le point F (représentant les quelques 5% de femmes que compte notre échantillon) est dans le quadrant F1 < 0, F2 > 0 : il s'agit donc de sujets ayant utilisé le tabac modérément et pendant une durée plutôt brève. Tout à l'opposé le point MAR4 (a changé trois fois de marques), est dans le quadrant F1 > 0, F2 < 0 : grands fumeurs, et fumeurs âgés.

2.2 Un sous-tableau de Burt avec 7 diagnostics : Les sept diagnostics énumérés au § 1.1 (y compris TEM : témoin indemne), sont croisés avec les modalités de variables qui sont celles de l'analyse précédente aux quelques modifications suivantes près.

La variable de sexe H-F a disparu, puisque (cf § 1.2) la présente analyse ne concerne que des hommes. Et on a figuré trois variables supplémentaires intéressantes :

Mode d'inhalation : LEGE , MODE (modéré) ; PROF (profond) : qui constituent une subdivision de la modalité INAL de la variable "Inhalation".

Cigarette bouche : CB01 (garder la cigarette tout le temps, même sans tirer de bouffées) ; CB02 (la plupart du temps) ; CB03 (≈ la moitié du temps).

Nombre de cigarettes total fumées ; pour avoir un indice de ce total on a multiplié le nombre de cigarettes fumées par jour, par le nombre d'années que le sujet a fumé ; et défini trois modalités :

NC01 : indice <450 ; NC02 : indice de 450 à 900 ; NC03 : indice >900.

Considérons le plan 1 × 2 issu de cette analyse. Sur l'axe 1, le point TEM (témoins) s'oppose à l'ensemble des cancers, notamment au point KEPI (cancer épidermoïde) qui définit la modalité de diagnostic la plus fréquente. Comme on pouvait l'attendre, vont avec TEM les modalités caractérisant un usage modéré et relativement peu prolongé du tabac. Ce dernier fait est remarquable, car les sujets étant appariés à âge égal, la distribution des âges des cancéreux et des témoins est la même. Les différences d'âge (de A >70 à A <49) s'inscrivent sur l'axe 2 : les sujets jeunes se distinguant par la fréquence du diagnostic d'adénocarcinome (ADEN ; le point KMIX, cancer mixte, qui représente 3 cas n'offre pas matière à interprétation).

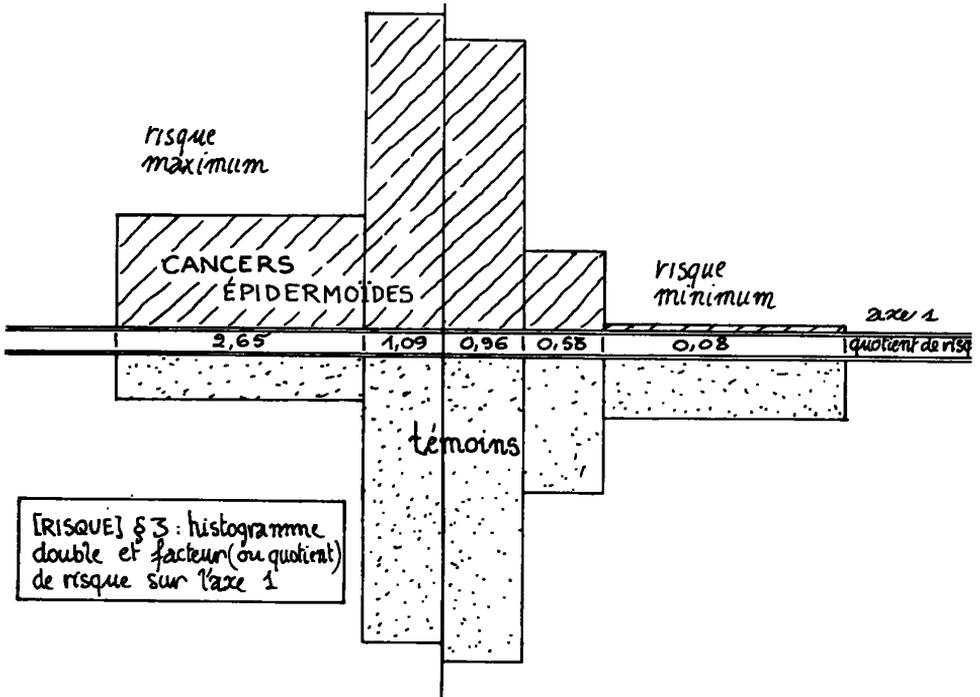
Ainsi qu'on l'a dit au § 1, il est utile de considérer la place des sujets (cancéreux et témoins) adjoints au sous-tableau de Burt en ligne supplémentaire : nous y reviendrons au § 3.

2.3 Un sous-tableau de Burt avec 5 diagnostics : Cette analyse ne concerne que les sujets masculins atteints de cancers bronchiques (à l'exclusion des témoins). Le diagnostic de cancer mixte très rare, a été éliminé. Les variables décrivant l'usage du tabac, ne diffèrent pas de celles de la présente analyse.

L'axe 1 issu de cette analyse est dominé par l'âge : quant aux diagnostics, le cancer épidermoïde, qui se déclare sur des sujets âgés, s'oppose à tous les autres diagnostics, particulièrement ADEN (adénocarcinome) et ANAP (anaplasique à petites cellules). La distinction entre ADEN et ANAP se fait sur l'axe 2 : on voit que l'adénocarcinome ne correspond pas à une intoxication tabagique importante ; alors que l'anaplasique, a eu une période d'intoxication relativement courte (de 25 à 34 années) mais à forte consommation.

3 Histogrammes et facteurs de risque

L'analyse du § 2.2, a montré sur le premier axe une nette opposition entre témoins et cancers ; avec une incidence minime de l'âge. Ceci incite à considérer le facteur 1 comme un facteur de risque assez pur, et à calculer des coefficients de risques d'après des histogrammes ainsi qu'on l'a proposé au § 0.1. De façon précise, on a tracé les histogrammes du facteur F1 d'une part pour la population des témoins, d'autre part pour la sous-population, suffisamment nombreuse des cancers épidermoïdes.



Ainsi qu'on l'a dit au § 1.2, le risque de cancer bronchique apparaît environ 20 fois plus élevé chez les fumeurs que chez les non-fumeurs. Au sein même des fumeurs, l'abscisse sur l'axe 1 permet de définir un facteur de correction au coefficient de risque moyen. Compte tenu des fluctuations d'échantillonnage, il nous a paru prudent de nous borner à des histogrammes à cinq créneaux. D'une extrémité à l'autre de l'axe (i.e. au créneau le plus à droite au créneau le plus à gauche) le coefficient de correction au risque moyen affectant les fumeurs passe de 0,08 à 2,65 donc le risque de cancer se trouve multiplié par $(2,65/0,08) \# 33$.

L'importance des variations ainsi mises en évidence, nous encourage à poursuivre l'étude des facteurs de risque par l'analyse multidimensionnelle.