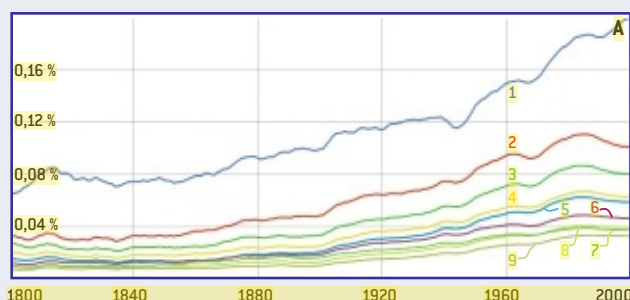
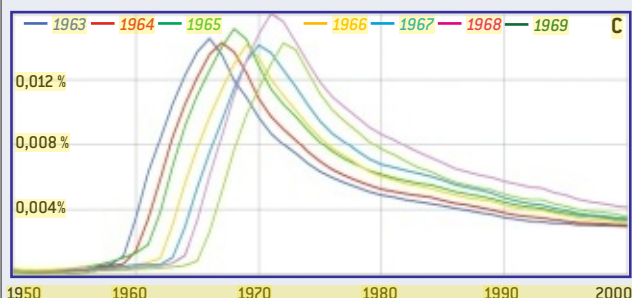
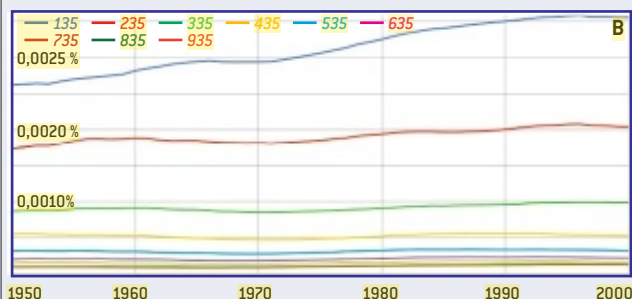


## 1. Des raffinements insoupçonnés de la loi de Benford

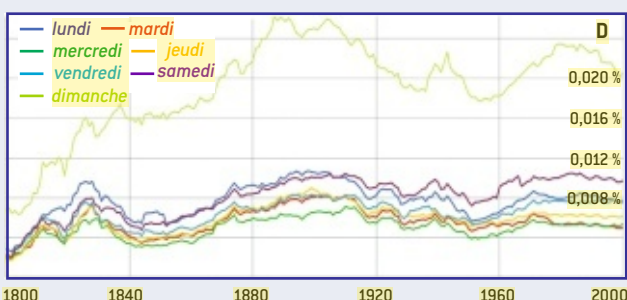
La loi de Benford est étonnante : si nous prenons des nombres au hasard dans une série en comportant beaucoup, comme les populations des villes, les longueurs des fleuves, les surfaces des pays, les cours de la Bourse, etc., les nombres commençant par le chiffre 1 sont plus nombreux que les nombres commençant par le chiffre 2, eux-mêmes plus nombreux que ceux commençant par 3, etc. Plusieurs tentatives d'explication du phénomène ont été proposées, pas toujours convaincantes.



Une façon de tester cette loi est de comparer les fréquences d'usages des signes isolés 1, 2, 3, 4, 5, 6, 7, 8, 9, ce que permet le système des *n-grammes* (groupes de *n* signes) associé à la base de cinq millions de livres. La loi est bien vérifiée. La pente montante de toutes les courbes provient sans doute de l'augmentation de la proportion des livres techniques dans l'ensemble des livres publiés (A). Plus frappant peut-être, la loi de Benford implique que 135 doit être plus souvent utilisé que 235, lui-même plus utilisé que 335, etc. C'est vrai (B).



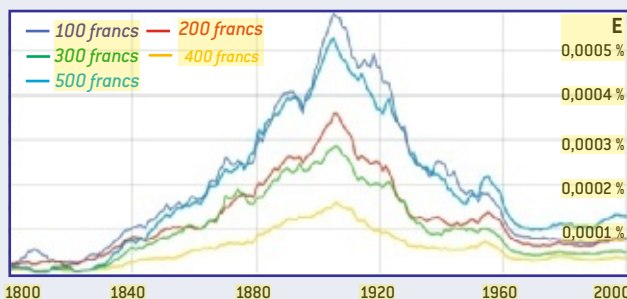
Bien sûr, des biais d'une autre nature se produisent par endroits, favorisant un nombre particulier et perturbant la loi de Benford générale au moins pendant quelques années. Les nombres ronds sont favorisés : 15 est plus utilisé que 14 ou 16, et 150 est aussi plus utilisé que 140 ou 160. Sans surprise encore, quand nous nous approchons de 1965, le nombre 1965 est de



plus en plus cité. Du coup, c'est trois ou quatre ans après 1965 que 1965 atteint son pic d'utilisation. Ce décalage temporel n'est pas l'explication de tout ce qu'on observe et, par exemple, l'année 1968 est celle qui va le plus haut parmi les années comprises entre 1963 et 1969 dans les livres en français (C), alors qu'un examen similaire montrerait que ce n'est pas le cas en anglais : les raffinements de la loi de Benford semblent sans limites.

Dans un ordre d'idées proche, l'étude de la fréquence d'utilisation des noms des jours de la semaine en français présente une structure imprévue, mais pas inexplicable (D). Dimanche domine ; jusqu'en 1910 environ lundi est en deuxième position, mais il se fait dépasser par samedi aujourd'hui bien installé en second. Viennent alors vendredi et jeudi, puis mardi et mercredi qui se disputent la dernière place.

Les courbes pour 100 francs, 200 francs, 300 francs, 400 francs et 500 francs présentent toutes un étonnant maximum vers 1910 (E). Pourquoi ?



Autre mystère : est-ce parce que nous devenons plus riches, est-ce un effet de l'inflation, ou alors faut-il chercher d'autres causes encore pour expliquer que (dans la course entre dix, cents, mille, millions, milliards) milliard vient en 1980 de passer en tête devant million qui lui-même avait battu mille vers 1900 (F) ?

