STATISTIQUE DESCRIPTIVE

Quelques extraits de la publication *Insee Première* (n° 1384 – décembre 2011[1]).

Distribution des salaires mensuels en 2009 et évolution entre 2008 et 2009 en euros constants

Déciles	Ensemble		Hommes		Femmes	
	2009	Évol.	2009	Évol.	2009	Évol.
D1	1 127	2,0	1 182	2,0	1 081	2,2
D2	1 256	2,1	1 326	2,1	1 184	2,1
D3	1 373	2,1	1 457	2,0	1 278	2,3
D4	1 500	2,0	1 593	1,9	1 378	2,3
D5 ou Médiane	**1 646**	**2,0**	**1 750**	**1,9**	**1 499**	**2,3**
D6	1 829	1,9	1 950	1,9	1 654	2,3
D7	2 075	2,0	2 228	1,9	1 867	2,3
D8	2 466	2,0	2 687	1,8	2 168	2,3
D9	3 255	1,5	3 596	1,1	2 751	2,3
D95	4 202	0,7	4 715	0,3	3 404	1,8
D99	7 499	–1,1	8 624	–1,4	5 472	0,8
Moyenne	**2 042**	**1,2**	**2 221**	**0,9**	**1 778**	**1,9**

Lecture : en 2009, 10 % des salariés en EQTP du secteur privé et semi-public gagnent un salaire mensuel net inférieur à 1 127 euros.
Champ : salariés en EQTP du secteur privé et semi-public, France.
Source : Insee, DADS.

« **En 2009, 10 % des salariés** (1er décile ou D1) **ont un salaire net mensuel en EQTP**[2] (équivalent temps plein) **inférieur à 1 127 euros**. En haut de l'échelle, **10 %** (9e décile ou D9) **disposent de plus de 3 255 euros** et les 1 % les mieux rémunérés (ou 99e centile) bénéficient de plus de 7 499 euros. L'évolution des salaires nets en euros constants a été positive pour l'ensemble de la hiérarchie salariale : entre 1,9 % et 2,1 % pour les huit premiers déciles et un peu plus faible pour le dernier décile (+ 1,5 %). Il s'agit de la poursuite d'une évolution observée depuis 2004 (+ 1,1 % en moyenne annuelle en euros constants pour le D1 contre + 1 % pour la médiane et + 0,8 % pour le D9). C'est le 9e décile qui augmente le moins sur ces deux années de crise. Mais ces mouvements relatifs demeurent d'ampleur modeste : le rapport entre les salaires des 1er et 9e déciles, un indicateur qui fournit une mesure de leur dispersion, est stable à 2,9 depuis 2004. **C'est au niveau des très hauts salaires que l'on mesure des évolutions plus différenciées** : le 99e centile a augmenté sensiblement plus vite que la médiane en 2007 (+ 2,4 %), puis encore de + 1 % en 2008, mais il baisse de 1,1 % en 2009. »

« ***L'écart salarial entre hommes et femmes demeure***

En 2009, **le salaire moyen des femmes progresse plus que celui des hommes** (+ 1,9 % en euros constants contre + 0,9 %). Ce constat se vérifie sur l'ensemble de la hiérarchie salariale mais, en particulier, pour les salaires les plus élevés, celui du 9e décile augmentant de 2,3 % pour les femmes contre 1,1 % pour les hommes. En effet, **c'est surtout parmi les cadres que se fait la différence** : le salaire moyen diminue de 2,2 % pour les hommes et augmente de 0,3 % pour les femmes entre 2008 et 2009. C'est dans les secteurs financiers que les salaires des cadres masculins ont le plus baissé : leur salaire moyen recule de 6,8 % en euros constants alors que celui des cadres féminins diminue de 0,7 % (10 % des cadres masculins et 13 % des cadres féminins travaillant dans les secteurs financiers). **Même s'il se réduit légèrement, l'écart salarial moyen entre hommes et femmes demeure important** (graphique) : **une salariée gagne, en moyenne en EQTP, 19,9 % de moins que son homologue masculin, contre 20,8 % en 2008**. C'est encore plus le cas chez les cadres où l'écart entre hommes et femmes est de 23,4 % bien qu'il se soit réduit plus fortement en 2009. Le salaire des hommes reste davantage dispersé que celui des femmes, les rapports interdécile valant respectivement 3,0 et 2,5. »

Salaires mensuels moyens et répartition des effectifs en EQTP

	Salaires bruts			Salaires net de tous prélèvements			Répartition des effectifs (%)	
	Euros courants		Euros constants	Euros courants		Euros constants		
	2008	2009	Évolution (%)	2008	2009	Évolution (%)	2008	2009
Ensemble	**2 682**	**2 708**	**0,9**	**2 016**	**2 042**	**1,2**	**100,0**	**100,0**
Cadres[1]	5 261	5 186	– 1,5	3 909	3 851	– 1,6	16,7	17,1
Prof. interm.	2 807	2 799	– 0,4	2 102	2 104	0,0	19,2	20,5
Employés	1 930	1 946	0,7	1 463	1 481	1,1	29,8	29,4
Ouvriers	2 009	2 042	1,5	1 529	1 563	2,1	34,2	32,9
Hommes	**2 917**	**2 938**	**0,6**	**2 199**	**2 221**	**0,9**	**100,0**	**100,0**
Femmes	**2 330**	**2 370**	**1,6**	**1 742**	**1 778**	**1,9**	**100,0**	**100,0**
Smic[2] (151,67h)	*1 305*	*1 329*	*1,7*	*1 025*	*1 044*	*1,8*	-	-

1. Y compris chefs d'entreprise salariés.
2. Smic en moyenne annuelle sur l'année civile.
Champ : salariés en EQTP du secteur privé et semi-public, France.
Source : Insee, DADS.

Salaires horaires moyens et répartition des effectifs[1]

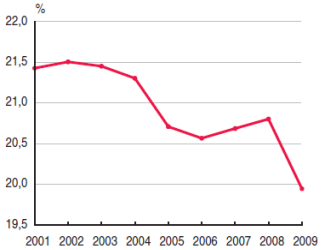
	Salaires bruts			Salaires nets de tous prélèvements			Répartition des effectifs (%)	
	Euros courants		Euros constants	Euros courants		Euros constants		
	2008	2009	Évolution (%)	2008	2009	Évolution (%)	2008	2009
Salariés à temps complet								
Hommes	18,97	19,13	0,8	14,29	14,45	1,0	65,2	64,9
Femmes	15,83	16,10	1,6	11,83	12,07	1,9	34,8	35,1
Ensemble	**17,87**	**18,07**	**1,0**	**13,43**	**13,62**	**1,3**	**100,0**	**100,0**
Cadres[2]	34,15	33,63	– 1,6	25,38	24,97	– 1,7	17,7	18,1
Prof. interm.	18,38	18,35	– 0,3	13,76	13,79	0,2	19,9	21,2
Employés	12,93	13,05	0,8	9,81	9,94	1,2	26,4	26,0
Ouvriers	13,21	13,49	2,0	10,04	10,31	2,6	36,0	34,7
Salariés à temps non complet								
Hommes	16,91	17,02	0,6	12,88	13,00	0,8	30,1	29,7
Femmes	13,80	13,98	1,2	10,33	10,52	1,8	69,9	70,3
Ensemble	**14,74**	**14,88**	**0,9**	**11,10**	**11,26**	**1,3**	**100,0**	**100,0**
Cadres[2]	29,45	29,84	1,2	21,81	22,15	1,5	12,0	12,6
Prof. interm.	17,46	17,24	– 1,4	13,14	13,00	– 1,2	15,0	16,1
Employés	11,64	11,70	0,5	8,81	8,90	0,9	51,0	50,1
Ouvriers	11,83	11,87	0,2	9,09	9,16	0,7	22,1	21,3
Smic	*8,61*	*8,77*	*1,8*	*6,76*	*6,88*	*1,7*	-	-

1. Effectifs en nombre d'heures travaillées.
2. Y compris chefs d'entreprise salariés.
Champ : salariés du secteur privé et semi-public, France.
Source : Insee, DADS.

Remarque de maître M :

lorsqu'on étudie les variations d'une quantité entre deux dates, cette quantité étant mesurée au moyen de sa valeur monétaire, ces variations sont perturbées par l'inflation qui a eu lieu entre ces deux dates, c'est-à-dire la diminution de la valeur de la monnaie. Si on ne corrige pas l'impact de l'inflation, on mesure ***à prix courants***. Si on corrige l'impact de l'inflation, on mesure l'évolution de cette quantité ***à prix constants***. *Source* : Wikipedia.

Écart entre le salaire moyen des hommes et celui des femmes (en EQTP)



Champ : salariés du secteur privé et semi-public, France.
Source : Insee, DADS.

« Les salaires annuels et les effectifs sont connus grâce aux DADS (Déclarations Annuelles de Données Sociales) que les entreprises adressent à l'administration. Tous les salariés, présents ou non toute l'année, sont concernés, à l'exception des personnels des services domestiques et des agents de la Fonction publique d'État. »

1 http://www.insee.fr/fr/ffc/ipweb/ip1384/ip1384.pdf

2 « Le salaire en équivalent temps plein (EQTP) est calculé en prenant en compte tous les postes de travail des salariés (y compris les postes à temps partiel). Chaque poste est pris en compte au prorata de son volume horaire de travail rapporté à celui d'un poste à temps complet. »

Table des matières

I. Définir et représenter une série statistique

Définitions :

• La ***population*** d'une série statistique est l'ensemble des éléments appelés « individus » sur lesquels porte l'étude statistique.

• Le ***caractère*** d'une série statistique est la propriété étudiée sur chaque individu. Il est dit :

- ..... lorsqu'il ne se traduit pas par une grandeur mesurable ;
- ..... lorsqu'il ne peut prendre qu'un nombre fini de valeurs numériques ;
- ..... lorsqu'il peut prendre une infinité de valeurs numériques.

Exemples :

Situation étudiée	Population	Caractère	Valeurs possibles du caractère	Type du caractère
Les notes du DS1 pour cette classe	Les élèves de cette classe	La note obtenue au DS1	0 ; 0,5 ; 1 ; 1,5 ; … jusqu'à 20	
La nationalité des albigeois	Tous les albigeois	La nationalité	Française, espagnole, anglaise, …	
Les tailles (en cm) des élèves du lycée	Les élèves du lycée	Les tailles (en cm)	Les nombres réels compris entre 1 et 240	

Définitions :

• L'***effectif*** d'une valeur du caractère est le nombre de fois où cette valeur apparaît dans la série.

• La ***fréquence*** d'une valeur est le quotient de l'effectif de cette valeur par l'effectif total.

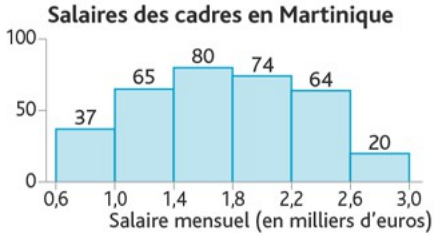
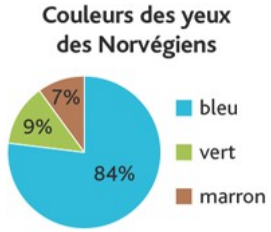
Exemple : si le caractère étudié est « les notes des 27 élèves de 2^{nde} au DS1 de mathématiques » et si 7 élèves ont eu 12 sur 20, alors 7 est l'effectif correspondant à la valeur 12 du caractère. La fréquence de la valeur 12 est alors : ≈ , ce qui signifie qu'environ … % des élèves ont obtenu 12 sur 20.

Propriété : La somme de toutes les fréquences est toujours égale à …

DÉMONSTRATION : admise, mais pas bien compliquée pour des padawans aussi intelligents que vous. Allez, demandez à votre maître, il sera content de l'intérêt que vous portez à sa discipline, et ainsi saura se montrer agréable pendant quelques minutes...

REPRÉSENTATIONS D'UNE SÉRIE STATISTIQUE

Selon le type du caractère, on utilise différentes représentations graphiques :

Caractères quantitatif discret ou qualitatif	*Caractère quantitatif continu*	*Tout type de caractère*
		

Remarque : pour un caractère quantitatif continu, les valeurs sont regroupés dans des intervalles appelés « classes ». Si ces classes n'ont pas toutes la même amplitude, alors on doit construire un histogramme « à pas non constant ». On décide que l'aire d'un rectangle est proportionnelle à l'effectif de la classe qu'il représente... Pour ceux que ça intéresse, et ceux que ça n'intéresse pas aussi, vous aurez ça en DM bientôt...

II. Indicateurs d'une série statistique

Une série statistique peut contenir de très nombreuses donnés (parfois plusieurs milliers).
Il est donc nécessaire de trouver une façon de résumer ces données.
D'après des donnés de l'Insee, *en 2008 un salarié français à temps complet du secteur privé et semi-public gagnait en moyenne 2 068 euros par mois, mais 10 % de ces salariés français à temps complet gagnait un salaire net mensuel inférieur à 1 124 euros*. On peut lire que *la moitié des salariés a touché moins de 1 653 euros nets*. Ces éléments sont appelés des « indicateurs ».
On différencie les ***indicateurs de position*** et les ***indicateurs de dispersion***.

II.1 Indicateurs de position

II.1.1. La moyenne

C'est l'indicateur le plus répandu. On compare souvent une note à la moyenne de la classe. En réalité, il existe plusieurs types de moyenne : on parlera ici seulement de *moyenne arithmétique*.

Définition : on considère une série statistique dont les valeurs du caractère sont $x_1, x_2, x_3, \ldots, x_p$ et les effectifs associés : $n_1, n_2, n_3, \ldots, n_p$.

La ***moyenne*** de cette série statistique, notée $\bar{x}$, a pour valeur : $\bar{x}=\dfrac{n_1\times x_1+n_2\times x_2+\ldots+n_p\times x_p}{n_1+n_2+\ldots+n_p}$.

Exemple « simplifié » : un 12 et deux 14 donnent une moyenne de $\dfrac{1\times 12+2\times 14}{1+2}=\dfrac{40}{3}\approx 13{,}33$.

Remarque : si on note f_i la fréquence de la valeur x_i , alors $\bar{x}=f_1x_1+f_2x_2+\ldots+f_px_p$.
Par exemple, si 2 personnes ont 16 ans alors que 8 autres ont 15 ans, pour obtenir l'âge moyen du groupe on peut calculer ..... au lieu de faire ..... .

II.1.2. La médiane

La médiane correspond à une valeur qui partage en deux parties (presque) égales la série statistique.

> **Définition :** une ***médiane*** d'une série statistique est un nombre, noté ***Me***, tel que :
> • au moins ..... des individus ont une valeur du caractère inférieure ou égale à ***Me*** ;
> • au moins ..... des individus ont une valeur supérieure ou égale à ***Me***.

En pratique, si on étudie un caractère quantitatif discret, en admettant que la série des données est ordonnée par ordre croissant, voici comment on détermine la médiane :
• Si la série est de taille impaire ($2n+1$), la médiane est la valeur du terme de rang …
• Si la série est de taille paire ($2n$), la médiane est …

II.1.3. Différence entre moyenne et médiane

⚠ Deux séries peuvent avoir la même moyenne mais des médianes très différentes.

Exemple : on considère les séries de notes suivantes :
série 1 : 03 ; 04 ; 05 ; 08 ; 10
série 2 : 02 ; 02 ; 02 ; 03 ; 07 ; 20.
La médiane de la série 1 est … . Sa moyenne est $(3+4+5+8+10)\div 5=30\div 5=6$.
La médiane de la série 2 est … . Sa moyenne est $(2+2+2+3+7+20)\div 6=36\div 6=6$.

⚠ Deux séries peuvent avoir la même médiane mais des moyennes très différentes.

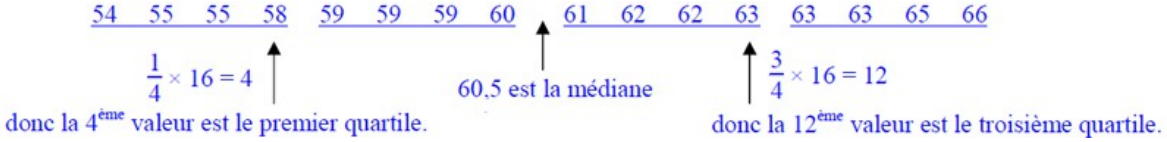
Exemple : on considère la série de 9 notes suivante :
02 ; 04 ; 05 ; 08 ; 10 ; 12 ; 12 ; 15 ; 16.
La médiane est 10 (la 5ème valeur) ; la moyenne est $(2+4+5+8+10+12+12+15+16)\div 9=84\div 9\approx$ **9,33** .
Si on rajoute les deux notes extrêmes 01 et 20, on obtient la série suivante :
01 ; 02 ; 04 ; 05 ; 08 ; 10 ; 12 ; 12 ; 15 ; 16 ; **20**
La médiane est toujours 10, mais la moyenne est : $(1+2+4+5+8+10+12+12+15+16+20)\div 11=105\div 11\approx$ **9,55** .

II.1.4. Les quartiles

> **Définitions :**
> • Le ***premier quartile*** de la série statistique, noté Q_1 , est la plus petite valeur telle qu'au moins … % des données soient inférieures ou égales à cette valeur.
> • Le ***troisième quartile*** de la série statistique, noté Q_3 , est la plus petite valeur telle qu'au moins … % des données soient inférieures ou égales à cette valeur.

En pratique, pour trouver le premier quartile Q_1 d'une série statistique rangée dans l'ordre croissant, on détermine le premier entier supérieur ou égal à $N\div 4$: cet entier est le rang de Q_1 .
Pour trouver le premier quartile Q_3 d'une série statistique rangée dans l'ordre croissant, on détermine le premier entier supérieur ou égal à $N\times 3\div 4$: cet entier est le rang de Q_3 .

Exemple : on a rangé par ordre croissant le nombre de pulsations cardiaques par minute de 16 élèves au repos :



• $Q_1=58$: un quart des élèves a 58 pulsations cardiaques par minute ou moins au repos.
• La médiane est 60,5 : la moitié des élèves ont moins de 60,5 pulsations cardiaques par minute au repos.
• $Q_3=63$: trois quarts des élèves ont 63 pulsations cardiaques par minute ou moins au repos.

II.2 Indicateurs de dispersion

Définitions :

• l'***étendue*** d'une série statistique est la différence entre la plus grande et la plus petite valeur de cette série ;

• l'***intervalle interquartile*** est l'intervalle $[Q_1 ; Q_3]$. L'écart $Q_3 - Q_1$ est appelé l'***écart interquartile***.

Exemple : pour la série statistique 3 ; 5 ; 6 ; 6 ; 8 ; 10 ; 12 ; 20 ; 20 ; 21 ; 23 ; 25 ; 27 :
L'étendue est …
Il y a 13 valeurs :

- $13 \div 4 = 3{,}25$ donc Q_1 est la … ème valeur, c'est-à-dire $Q_1 =$ …
- $13 \times 3 \div 4 = 9{,}75$ donc Q_3 est la … ème valeur, c'est-à-dire $Q_3 =$ …
- $13 = 2 \times 6 + 1$ donc la médiane Me est la … ème valeur : $Me =$ …

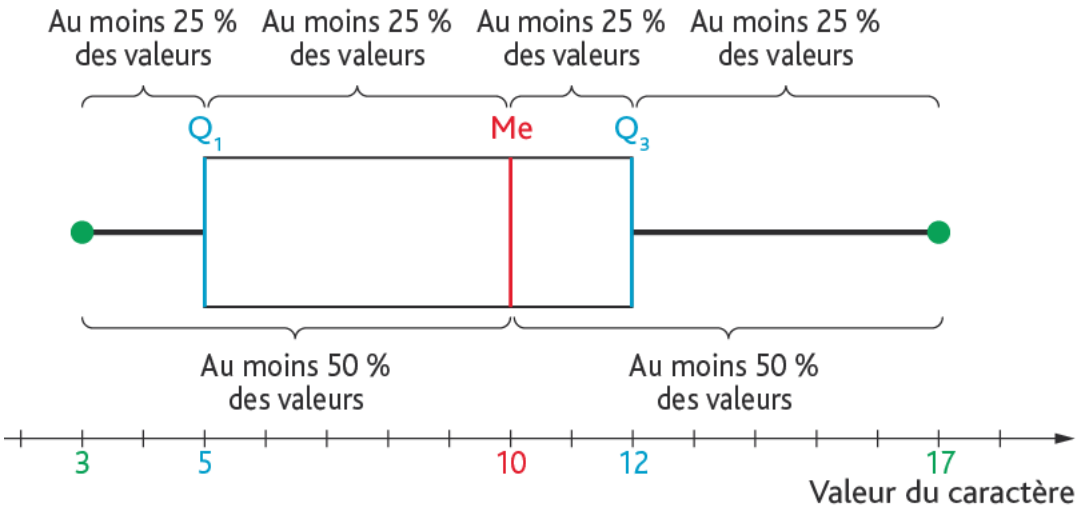
L'intervalle interquartile est donc l'intervalle $[6 ; 21]$ et l'écart interquartile est $21 - 6 = 15$.
Plutôt que de présenter cette série, on pourrait donner le tableau suivant :

Effectif total	Min.	Max.	Q_1	Me	Q_3
…	…	…	…	…	…

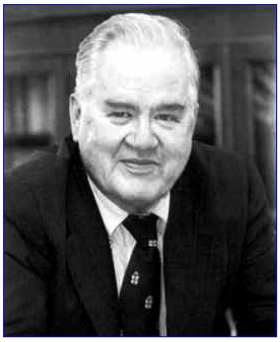
II.3 Résumé d'une série statistique

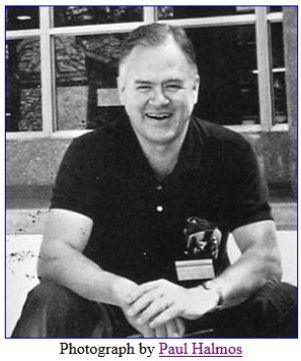
On peut résumer les indicateurs d'une série statistique par un **diagramme « en boîte »**, ou « diagramme de Tukey », ou « boîte à moustaches ».

Exemple pour la série 3 ; 3 ; 5 ; 6 ; 7 ; 10 ; 10 ; 11 ; 12 ; 12 ; 13 ; 17 :



III. Un peu de culture sur le mec aux moustaches




Photograph by Paul Halmos

John Wilder **Tukey** (16 juin 1915 - 26 juillet 2000) est l'un des plus importants statisticiens américains du XX^{e} siècle.

« Dans les années 1950, Tukey se penche sur le formidable outil pour les mathématiques et la statistique qu'apporte l'ordinateur et que sera l'informatique. On lui doit (1949) l'acronyme *bit* pour désigner le **BI**nary digi**T** (unité élémentaire d'information pouvant ne prendre que deux valeurs : 0 ou 1) évoqué cependant bien auparavant (1938) par le mathématicien Claude Shannon. ».

« Tukey serait aussi à l'origine du mot *software*, forgé (1947) sur et par opposition à *hardware* = quincaillerie (la machine), pour désigner ce que nous appelons logiciel en français (1972, *logic* = logique, *iel* = fin du mot matériel = *ware*). Bizarrement, le terme *hardware* fut conservé en France et accepté officiellement en 1974. »

« En 1965, il est à la tête du Département de statistiques de la célèbre université de Princeton. Les États-Unis surveillent alors le développement de l'armement nucléaire de l'URSS. Afin de localiser les explosions d'essais nucléaires soviétiques (par l'étude du spectre du signal engendré par les vibrations telluriques des explosions), on demande à Tukey une méthode de calcul rapide d'un outil mathématique (la transformée de Fourier discrète) afin d'utiliser efficacement l'outil informatique dans ses recherches. Avec son compatriote Cooley, Tukey présente son algorithme, appelé FFT (*Fast Fourier Transform* = Transformée de Fourier rapide) dont l'importance est aujourd'hui grande en théorie du signal (discipline qui développe et étudie les techniques de traitement, d'analyse et d'interprétation des signaux. Parmi les types d'opérations possibles sur ces signaux, on peut dénoter le contrôle, le filtrage, la compression de données, la transmission de données, le débruitage, la déconvolution [opération d'amélioration d'une image par un traitement numérique], la prédiction, etc). »
Ce travail aide par exemple les astronomes à déterminer le spectre de lumière venant d'une étoile, bien plus rapidement qu'auparavant...

« Il publia en 1977 un livre de référence, *Exploratory Data Analysis*, sur les méthodes d'analyse et de présentation des données. Il y présente le principe de la boîte à moustaches, qu'il a inventé. »

« Dans le cadre d'un programme confidentiel de recherche gouvernemental, il a probablement contribué au projet de développement de l'avion espion Lockheed U-21,2,3. »

« Il a inspiré, avec Lyman Spitzer Jr, la fabrication du télescope Hubble. »

Sources : *Wikipedia* et *http://serge.mehl.free.fr/chrono/Tukey.html*, entre autres.