Instrumentaliser les données ?

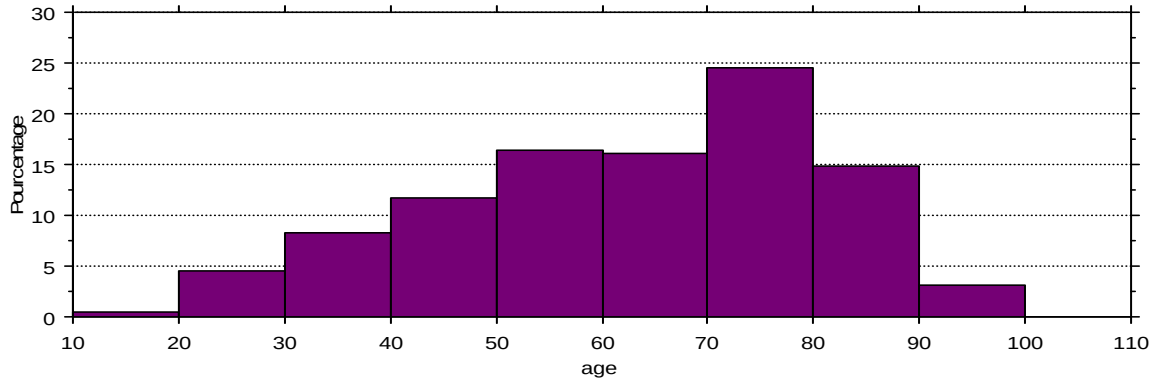
Maïeutique en statistique

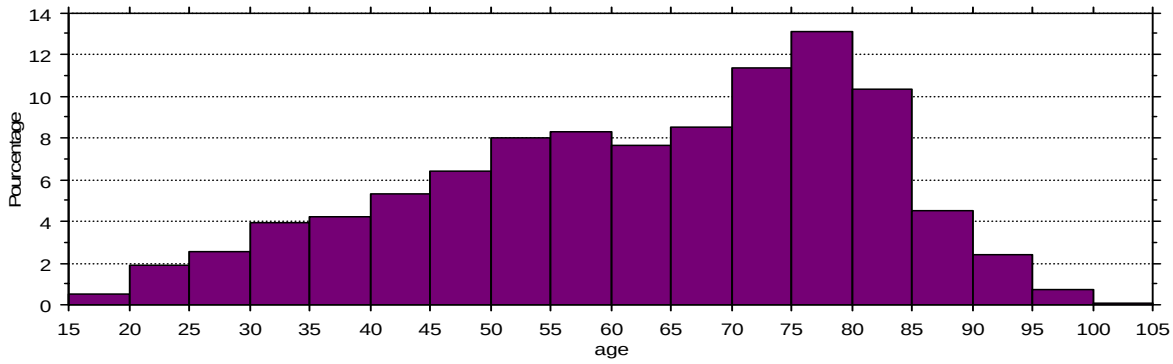
Léo Sila

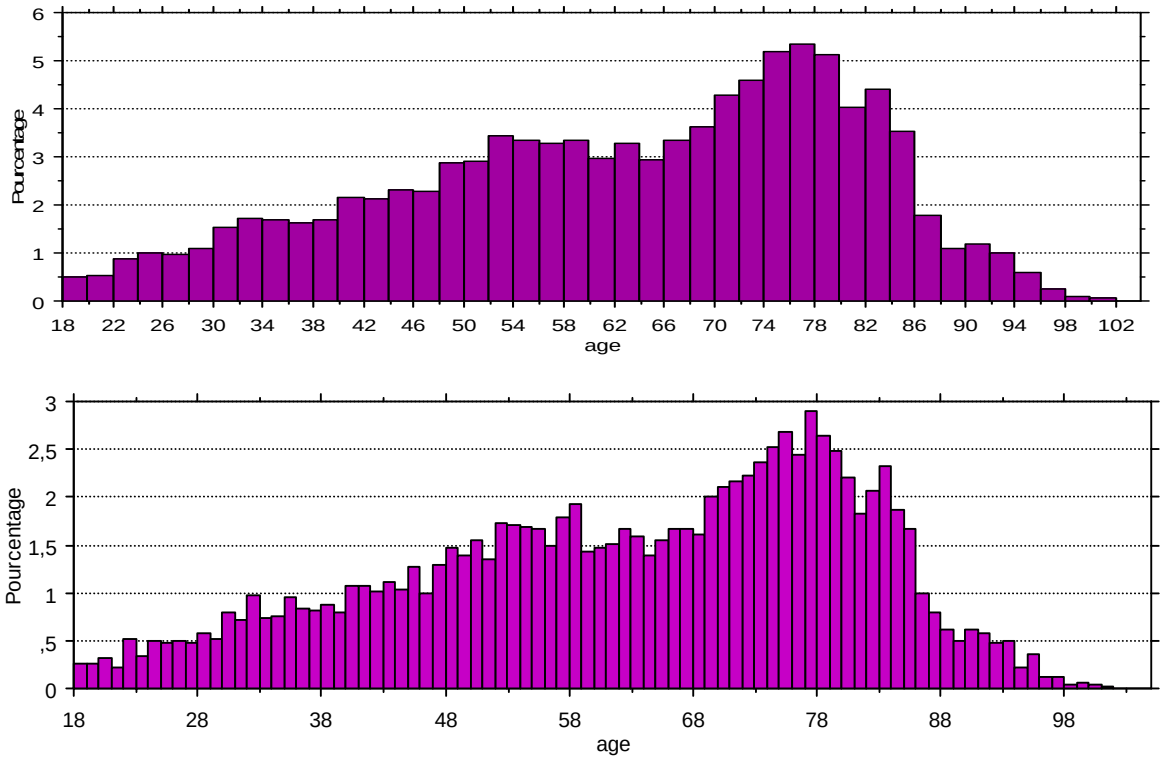
Pièce en trois actes

Acte I
De quoi s'agit-il ?

M. : J'ai tracé ces histogrammes (*il montre les histogrammes ci-dessous).* Qu'en dis-tu ?



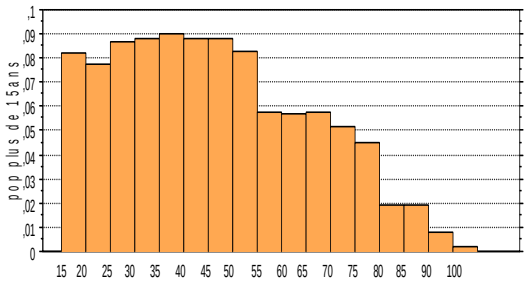


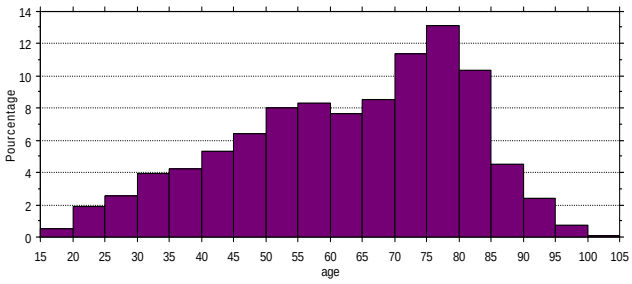


S. : Je dis : ce sont des histogrammes. En violet, comme ça, c'est joli.
M. : L'histogramme avec un pas de deux ans nous a paru le plus plaisant : on ne perd pas beaucoup de précisions, et il est assez régulier.
S. : Voici un choix plein de bon sens …
M. : Oh, j'ai oublié de te dire le principal : on a 7515 données. On ne peut donc pas les regarder une à une, contrairement aux points des matchs de foot[1]. On a suffisamment de données pour que tous les histogrammes produits se ressemblent, ce n'est pas comme pour les données du foot…
S. : Un histogramme est ici très efficace pour y voir un peu clair sur les données. Mais 7515 données, est-ce bien là *le principal* ?
M. : Je fais un cours sur les histogrammes, noter quelque part le nombre des données est donc important.
S. : Bien sûr. Mais de quelles données s'agit-il ?
M. : De l'âge de 7515 personnes. Je pourrais presque ne pas dire que ce sont des âges.
S. : Pour s'exercer à tracer des histogrammes avec des données dont on ne connaît même pas l'unité, 500 ou 50 0000 données fabriquées à ta guise feraient aussi l'affaire. Inutile d'avoir 7515 ou 8234 données d'expérience, si tu n'en dis pas un peu plus.
M. : Il s'agit ici des âges de personnes d'au moins 18 ans qui consultent des médecins de ville ou à l'hôpital pour des douleurs à la jambe : ça peut avoir des causes diverses, un excès de sport par exemple, mais aussi un caillot (un thrombus) dans une veine. Ces âges ont été recueillis dans le cadre d'une grande enquête sur des facteurs de risque de la maladie thrombo-embolique veineuse (M.T.E.). C'est un projet de recherche clinique, en cours de réalisation.
S. : Que disent les résultats de cette enquête ?

[1] Voir texte : Images inhabituelles du foot

M. : Les résultats sont en cours d'étude. Les données ne sont pas publiques, mais pour mon enseignement, on m'a autorisé à utiliser un sous-fichier avec l'âge, le sexe et une variable qui code la présence ou l'absence de M.T.E.[2]
S. : Donc il s'agit de personnes qui consultent pour des symptômes précis. Ce n'est pas du tout des gens de plus de 18 ans « choisis au hasard dans la population française » ! D'ailleurs, il n'y a qu'à regarder ; moi aussi, j'ai des histogrammes dans mes réserves ; voici, en orange, à gauche de l'histogramme de ta population, celui de la population des plus de 15ans, pour toute la France métropolitaine[3]. La population française est plus jeune.





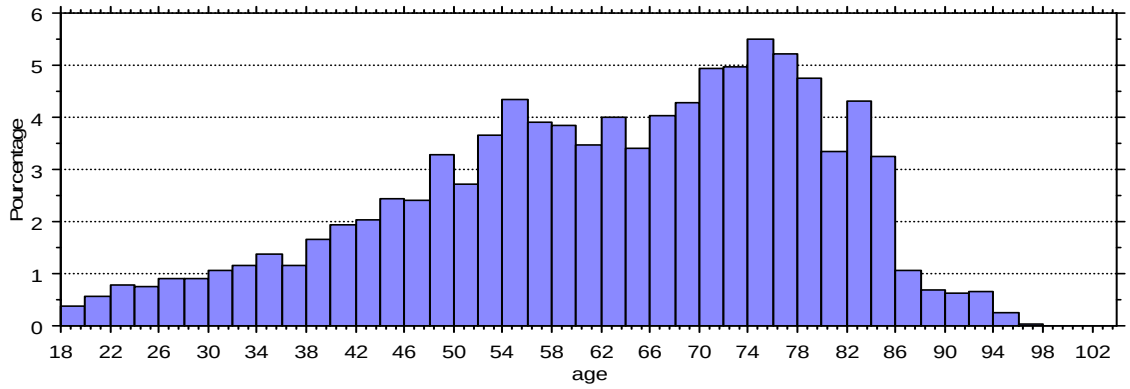
M. : Les jeunes vont moins chez le médecin, ce n'est pas un scoop.
S. : Tout fichier statistique digne de ce nom doit contenir un scoop ?
M. : Non, bien sûr ; cela n'empêche pas d'observer et résumer les données. L'âge moyen des 7515 patients est de 62,3 ans, l'âge le plus élevé est 101 ans, la médiane est de 65 ans, légèrement supérieure à la moyenne, ce qui est cohérent avec l'absence de symétrie des données autour de la moyenne.
S. : C'est toujours bien d'observer les données, pour la suite.
M. : La suite ?

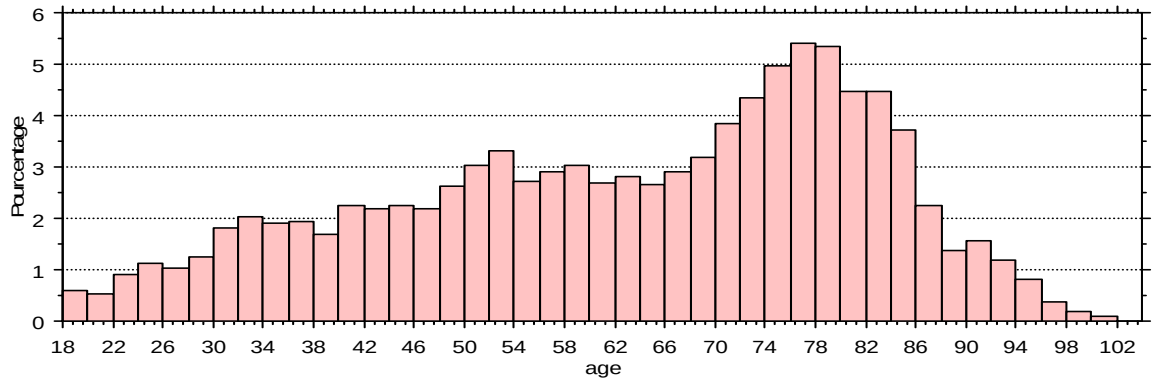
Acte II
Les âges des hommes et des femmes.

S. : Est-ce qu'il y a autant de femmes que d'hommes dans la population étudiée ?
M. : Ah, non. Il y a 38,8 % d'hommes et 61,2% de femmes.
S. : Parmi les personnes qui consultent pour des douleurs à la jambe, il y a ainsi presque 2/3 de femmes. C'est presque un scoop…
M. : Oh, ce qui tient du scoop pour certains est bien connu pour d'autres, voire inintéressant.
S. : Est-ce que les répartitions en âge des femmes et des hommes se ressemblent néanmoins.
M. : Je dois continuer mon cours, passer à autre chose, je ne peux pas plus m'attarder sur ces données.
S. : Tu aurais pu simuler des données : sans contexte, c'est nettement moins encombrant.

[2] On trouvera dans un fichier à part un fichier contenant les données traitées ici.
[3] Données issues de la pâge : http://www.recensement.insee.fr/pyramide_âges_france.html

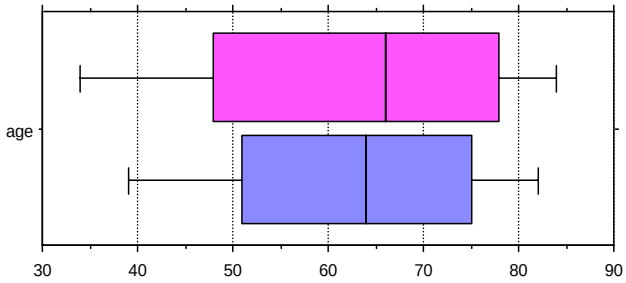
M. : Mais le contexte est intéressant ! Voila les répartitions *(il montre les dessins ci-dessous)* ; pour faciliter la lecture : garçons en bleu, filles en rose.





	Moy.	Dév. Std	Nombre	Minimum	Maximum	Etendue	Médiane
age	62,1	16,5	2919	18,0	96,0	78,0	64,0

	Moy.	Dév. Std	Nombre	Minimum	Maximum	Etendue	Médiane
age	62,5	18,8	4596	18,0	101,0	83,0	66,0



S. : Finalement, les proportions d'hommes et de femmes ne sont pas les mêmes, mais les répartitions en âge sont quasiment identiques. Encore un scoop pour qui le considère comme tel !
M. : Pourquoi dis-tu « finalement » ?
S : Je croyais que tu te servais de ces données pour parler d'histogramme, c'est tout.
M. : Mais ces données là, elles sont exceptionnelles ! Il a fallu des années de travail pour les recueillir. Rends-toi sur le site : http://www.optimev.net.
S. : Tu peux faire travailler des élèves de ta classe sur des sous fichiers de cas tirés au hasard dans ce gros fichiers : ils pourront observer la fluctuation d'échantillonnage au niveau des moyennes, des médianes, etc.
N. : Mais on est toujours dans l'instrumentalisation !
S. : Tu ne m'as pas parlé tout à l'heure de M.T.E. ?

Acte III
Maladie thrombo-embolique, âge et sexe

M. Je pensais que la maladie-thrombo-embolique était surtout une maladie du troisième ou quatrième âge. J'ai calculé les moyennes d'âge, pour les hommes et les femmes, selon qu'ils ont ou non une M.T.E (*il montre le tableau ci-dessous*). Je suis un peu perplexe : l'âge moyen des patients du fichier est 62,3 ans. L'âge moyen des hommes atteints de M.T.E. est 62,3ans celui des hommes sans M.T.E. est 62,0 ans ; pour les femmes, la moyenne d'âge de celles qui ont une M.T.E. est 64,5 ans contre 61,8 ans pour celles qui n'ont pas de M.T.E. : cette différence n'est pas bien grande !

	Moy.	Nombre	Minimum	Maximum	Etendue	Médiane
age, Total	62,3	7515	18,0	101,0	83,0	65,0
age, 1, 0	62,0	1936	18,0	96,0	78,0	64,0
age, 1, 1	62,3	983	19,0	95,0	76,0	65,0
age, 2, 0	61,8	3383	18,0	101,0	83,0	64,0
age, 2, 1	64,5	1213	18,0	99,0	81,0	69,0

S. : Les patients de ton fichier ne consultent-ils pas pour des douleurs à la jambe ou autre symptômes évocateurs de M.T.E. ?
M. : J'avais oublié ! Finalement, on a envie de dire que le symptôme « douleur à la jambe » n'est pas tellement plus évocateur d'une M.T.E. chez les jeunes que chez les seniors.
Pour me clarifier les idées, j'ai coupé la population entre deux tranches d'âge (*il montre le tableau ci-dessous*) : il y 23,5% de M.T.E. chez les moins de 30 ans, et 29,5% chez les plus de 30 ans, homme et femmes réunis. C'est donc un peu plus élevé dans le groupe des plus de 30 ans, mais la différence n'est pas si grande que ça.

	Sans M.T.E.	Avec M.T.E.	total
Âge ≤30	286 **76,5%**	88 **23 ,5%**	374 **100%**
Âge>30	5033 **70,5%**	2108 **29,5%**	7141 **100%**
total	5319 **70,8%**	2196 **29,2%**	7515 **100%**

S. : Pas si grande que ça…. c'est toi qui le dit ! C'est relatif, grand ou pas grand…. Si on fait le rapport des fréquences de la M.T.E. chez les plus de 30 ans et les moins de 30 ans, on trouve 1,25. Ce rapport des proportions, qui vaut ici 1,25, les médecins l'appellent le risque relatif ; parmi les gens qui ont mal à la jambe, le risque d'une M.T.E. est 1,25 fois plus grand dans le groupe *âgé* que dans le groupe *jeune*. Moi, je trouve que c'est beaucoup Mais au fait, pourquoi as-tu coupé à 30 ans ?

M : Tu ne me laisses même pas le temps de te montrer la suite. J'ai aussi découpé en deux classes avec un seuil à 50 ans, un autre à 70 ans (*il montre les tableaux ci-dessous*).

	Sans M.T.E.	Avec M.T.E.	total
Âge ≤30	1467 **73,5%**	524 **26 ,5%**	1991 **100%**
Âge>30	3852 **69,9%**	1672 **30,1%**	5524 **100%**
total	5319 **70,8%**	2196 **29,2%**	7515 **100%**

	Sans M.T.E.	Avec M.T.E.	total
Âge ≤30	3218 **71,9%**	1257 **28 ,1%**	4475 **100%**
Âge>30	2101 **69,1%**	939 **30,9%**	3040 **100%**
total	5319 **70,8%**	2196 **29,2%**	7515 **100%**

S. : Si on coupe à 50 ans, le risque relatif est RR=1,15.
Si on coupe à 70 ans, le risque relatif est RR=1,10. Le risque relatif décroît à mesure que le seuil de découpage s'élève. A-t-on une raison de choisir un seuil de découpage plus qu'un autre ?
N. : Tout dépend de ce qu'on veut faire. Par exemple, on aurait pu faire une étude chez les femmes, en coupant à 50 ans, qui est environ l'âge où elles arrêtent la contraception et ne peuvent plus être enceintes, ces deux éléments étant connus pour augmenter légèrement le risque de M.T.E.. Mais plus généralement, les médecins étudient le risque relatif ou un indice analogue, en considérant que c'est une fonction de l'âge, et de certains facteurs tels le sexe, les antécédents dans cette maladie et quelques autres facteurs.

S. : A propos, quand on passe d'un seuil à 30 ans à un seuil à 50 ans, puis à 70 ans, à chaque fois le risque de M.T.E. augmente dans chaque groupe. Je ne vois pas très comment c'est possible.
N. : J'espérais bien que tu allais me poser cette question ! Je vais prendre des données fictives et tu vas comprendre comment ça peut se passer :

Données fictives	**Âge ≤30**	**30<Âge≤50**	**50<Âge**
Avec M.T.E.	**10**	**20**	**40**
Sans M.T.E.	**90**	80	60

Avec un découpage en deux classes d'age : les moins de 30 ans et les plus de 30 ans, le risque chez les jeunes est 0,10 contre 0,30 chez les moins jeunes.
Avec un découpage en deux classes d'age : les moins de 50 ans et les plus de 50 ans, le risque chez les jeunes est 0,15 contre 0,40 chez les moins jeunes.

S. : Evidemment…Enfin, heureusement que le pourcentage de M.T.E. chez les 7515 patients est toujours compris entre celui des jeunes et celui des moins jeunes, ça on en est toujours sûr ! Et à propos, les hommes et les femmes qui ont une douleur à la jambe sont-ils égaux, face à cette maladie ?

N. : Voilà.

	Sans M.T.E.	**Avec M.T.E.**	**total**
Hommes	1936 **66,3%**	983 **33 ,7%**	2919 **100%**
femmes	3383 **73,6%**	1213 **26,4%**	4596 **100%**
total	5319 **70,8%**	2196 **29,2%**	7515 **100%**

S. : Donc un homme a ici 1,28 fois plus de chances d'avoir une M.T.E. qu'une femme.
N. : Oui, dans notre population qui n'est pas la population générale !

S. : Finalement, …..les données n'ont pas été « *instrumentalisées à 100%* ».