Fiche TD avec le logiciel R : convergencet

Étude empirique de l'approximation de la loi de Student par la loi de Laplace-Gauss

J.R. Lobry

Pourquoi à partir de n = 30 on considérait que l'on avait un grand échantillon ?

1 Historique

La question de la légitimité de l'utilisation de la loi de Laplace-Gauss comme approximation de la loi de Student a été clairement posée dans un article de 1908 de William Sealy Gosset [7], publié sous le pseudonyme de *Student*, d'où la loi éponyme.

> **Again, although it is well known that the method of using the normal curve is only trustworthy when the sample is "large," no one has yet told us very clearly where the limit between "large" and "small" samples is to be drawn.**
>
> **The aim of the present paper is to determine the point at which we may use the tables of the probability integral in judging of the significance of the mean of a series of experiments, and to furnish alternative tables for use when the number of experiments is too few.**

À l'époque où écrit William Sealy Gosset, le théorème-limite central de la théorie des probabilités est bien connu, Karl Pearson [3] en attribue la paternité à un article en latin d'Abraham de Moivre de 1733 dont on peut trouver un fac simile dans un article d'Archibald [1]. Mais il faut attendre 1920 pour que Pólya lui donne [6] son appellation moderne *der zentralen Grenzwertsatz*. À noter que l'adjectif *zentralen*, central au sens de très important, modifie le nom composé *Grenz-wert-satz*, littéralement limite-valeur- théorème c'est à dire un théorème-limite. Il n'y a pas d'ambiguïté en allemand : ce qui est central c'est le théorème, pas la limite. Une traduction moins mot-à-mot pourrait être par exemple : le théorème fondamental de convergence en loi de la théorie des probabilités. La dénomination anglaise *central limit theorem* n'est pas non plus très heureuse, retraduite en français cela donne parfois le théorème central limite qui sonne curieusement comme un théorème qui serait central ... mais un peu limite ! On rencontre aussi, sans doute par attraction parce qu'il est question de variables centrées, le théorème de la limite centrale, ou le théorème de la limite centrée. Le fameux et facétieux Edward Nelson parle dans son livre [5]

William Sealy Gosset (1876-1937). La politique d'embargo de la brasserie Guinness qui l'employait l'a empêché de publier ses travaux en son nom propre, il a utilisé comme nom de plume *Student*, l'étudiant. La raison de l'embargo est que la brasserie Guinness voulait garder secret le fait qu'elle employait des statisticiens pour augmenter sa compétitivité industrielle. (source : `http://www.bobabernethy.com/bios_stats.htm`).

La brasserie *Arthur Guinness Son & Co.* a été fondée en 1759 par Arthur Guinness à Dublin en Irlande. Elle a employé William Sealy Gosset à partir de 1899, en particulier pour sélectionner les variétés d'orge (*Hordeum vulgare*) les plus intéressantes d'un point de vue industriel. (source : `http://www.wikipedia.org`).

du théorème de de Moivre-Laplace-Lindeberg-Feller- Wiener-Lévy-Doob-Erdös-Kac-Donsker-Prokhorov.

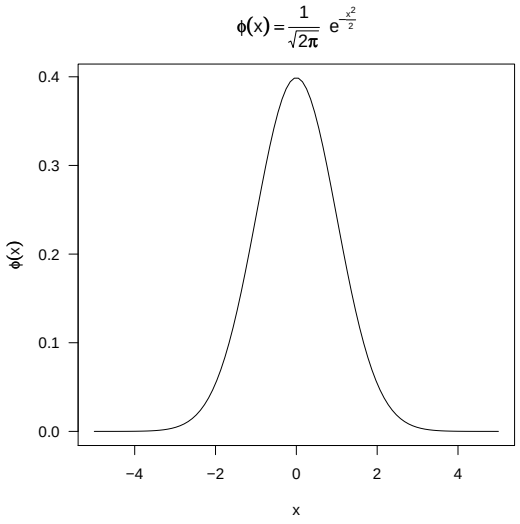
2 La loi de Laplace-Gauss comme approximation de la loi de Student

$$P\left(X \leq x\right) = F_N\left(x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

$$\mu = E\left(X\right) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} t e^{-\frac{t^2}{2}} dt = 0 \quad \sigma^2 = E\left(X^2\right) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} t^2 e^{-\frac{t^2}{2}} dt = 1$$

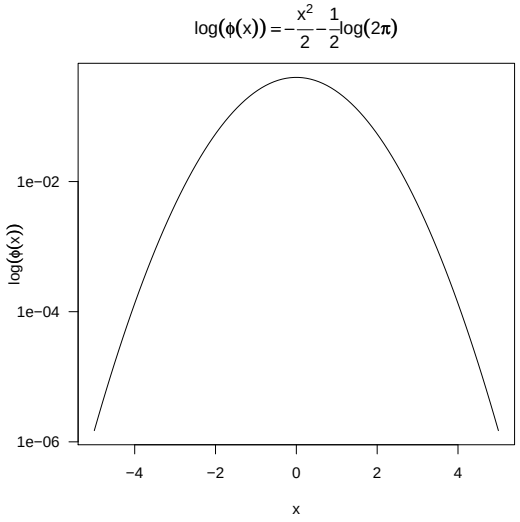
Représentation graphique de la fonction de densité de probabilité $\phi(x)$:

```
x <- seq(from = -5, to = +5, length = 100)
plot(x, dnorm(x), type = "l", las = 1, ylab = expression(phi(x)),
    main = expression(phi(x) == frac(1, sqrt(2 * pi)) * phantom(0) *
        e^-frac(x^2, 2)))
```

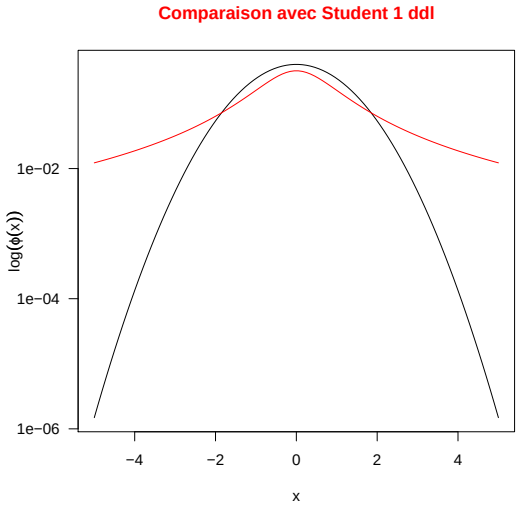


Les physiciens aiment bien passer en coordonnées semi-logarithmiques :

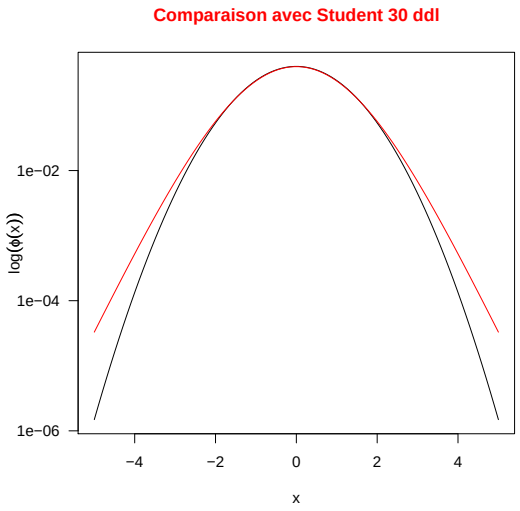
```
plot(x, dnorm(x), type = "l", las = 1, ylab = expression(log(phi(x))),
    main = expression(log(phi(x)) == -frac(x^2, 2) - frac(1, 2) *
        log(2 * pi)), log = "y")
```



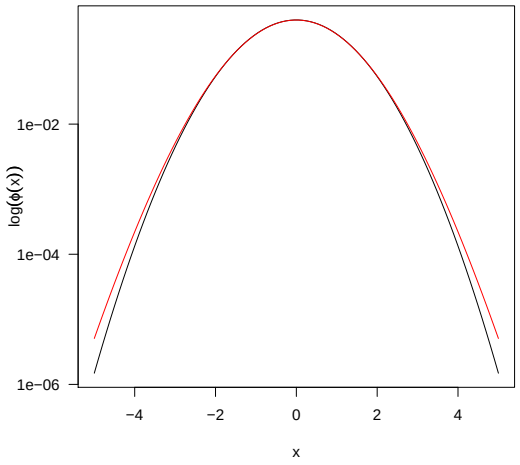

```
df <- 1
plot(x, dnorm(x), type = "l", las = 1, ylab = expression(log(phi(x))),
    main = paste("Comparaison avec Student", df, "ddl"), log = "y",
    col.main = "red")
lines(x, dt(x, df = df), col = "red")
```




```
df <- 30
plot(x, dnorm(x), type = "l", las = 1, ylab = expression(log(phi(x))),
    main = paste("Comparaison avec Student", df, "ddl"), log = "y",
    col.main = "red")
lines(x, dt(x, df = df), col = "red")
```



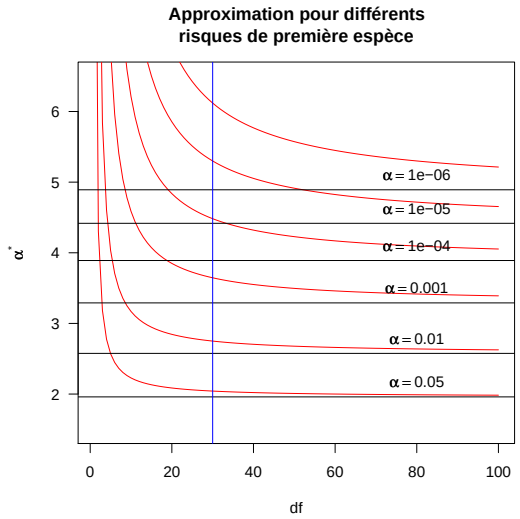

```
df <- 100
plot(x, dnorm(x), type = "l", las = 1, ylab = expression(log(phi(x))),
    main = paste("Comparaison avec Student", df, "ddl"), log = "y",
    col.main = "red")
lines(x, dt(x, df = df), col = "red")
```



La queue de la distribution est donc bien plus épaisse dans le cas de la loi de Student. C'est embêtant parce c'est la région utile pour nous en pratique. Voyons l'erreur que l'on commet pour différents risques de première espèce α en utilisant une loi normale à la place d'une loi de Student.

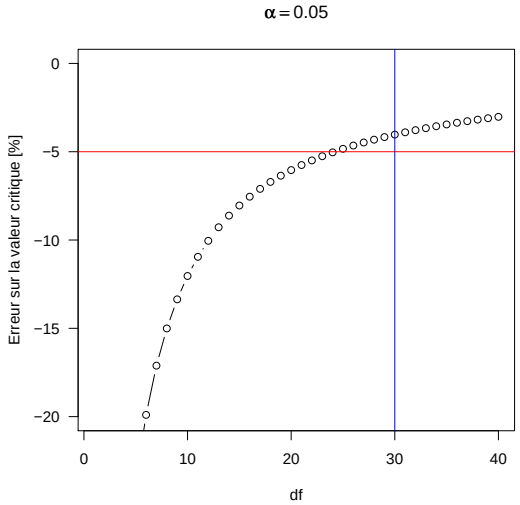
```
df <- 1:100
alpha <- 0.05
critique <- sapply(df, function(x) qt(1 - alpha/2, df = x))
plot(df, critique, type = "l", col = "red", ylim = c(1.5, 6.5),
    las = 1, main = "Approximation pour diff\351rents\nrisques de premi\350re esp\350ce",
    ylab = expression(alpha^{
        "*"
    }))
abline(h = qnorm(1 - alpha/2))
```

```
text(0.8 * max(df), qnorm(1 - alpha/2), labels = bquote(alpha ==
    .(alpha)), pos = 3)
for (alpha in 10^-(2:6)) {
    critique <- sapply(df, function(x) qt(1 - alpha/2, df = x))
    lines(df, critique, col = "red")
    abline(h = qnorm(1 - alpha/2))
    text(0.8 * max(df), qnorm(1 - alpha/2), labels = bquote(alpha ==
        .(alpha)), pos = 3)
}
abline(v = 30, col = "blue")
```



Voilà donc d'où vient le fameux $n = 30$ pour décréter que l'on a un grand échantillon. Quand on travaille avec un risque de première espèce $\alpha = 0.05$, on ne fait pas une grosse erreur : la valeur critique de la loi de Student est de 2.042272 pour 30 ddl, que l'on approxime par 1.959964 avec la loi normale, soit une différence de -4.03 %. Tiens, tiens, comment évolue l'erreur relative sur la valeur critique avec la taille de l'échantillon ?

```
df <- 1:40
alpha <- 0.05
errcritique <- sapply(df, function(x) 100 * (qnorm(1 - alpha/2) -
    qt(1 - alpha/2, df = x))/qt(1 - alpha/2, df = x))
plot(df, errcritique, type = "b", ylim = c(-20, 0), las = 1, ylab = "Erreur sur la valeur critique [%]",
    main = bquote(alpha == .(alpha)))
abline(h = -5, col = "red")
abline(v = 30, col = "blue")
```



J'imagine que ce qui a du se passer c'est que nos ancêtres qui ont tabulé laborieusement les tables de la loi de Student à la main ont estimé que cela ne valait plus la peine de se fatiguer à partir du moment où l'erreur sur la valeur critique était du même ordre de grandeur que celui du risque de première espèce. D'où $n = 30$ en arrondissant à la dizaine supérieure.

La première tabulation de la loi de Student publiée par William Gosset en 1908 [7] est reproduite ci-après :

SECTION VII. *Tables of* $\frac{n-2}{n-3}\,\frac{n-4}{n-5}\cdots\begin{pmatrix}\frac{3}{2}\cdot\frac{1}{2}\,n \text{ odd}\\ \frac{2}{1}\cdot\frac{1}{\pi}\,n \text{ even}\end{pmatrix}\int_{-\frac{\pi}{2}}^{\tan^{-1}x}\cos^{n-2}\theta\,d\theta$

for values of n *from* 4 *to* 10 *inclusive.*

Together with $\frac{\sqrt{7}}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{7x^2}{2}}dx$ *for comparison when* $n=10$.

$z\left(=\frac{x}{s}\right)$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$	$n=9$	$n=10$	For comparison $\left(\frac{\sqrt{7}}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{7x^2}{2}}dx\right)$
·1	·5633	·5745	·5841	·5928	·6006	·60787	·61462	·60411
·2	·6241	·6458	·6634	·6798	·6936	·70705	·71846	·70159
·3	·6804	·7096	·7340	·7549	·7733	·78961	·80423	·78641
·4	·7309	·7657	·7939	·8175	·8376	·85465	·86970	·85520
·5	·7749	·8131	·8428	·8667	·8863	·90251	·91609	·90691
·6	·8125	·8518	·8813	·9040	·9218	·93600	·94732	·94375
·7	·8440	·8830	·9109	·9314	·9468	·95851	·96747	·96799
·8	·8701	·9076	·9332	·9512	·9640	·97328	·98007	·98253
·9	·8915	·9269	·9498	·9652	·9756	·98279	·98780	·99137
1·0	·9092	·9419	·9622	·9751	·9834	·98890	·99252	·99820
1·1	·9236	·9537	·9714	·9821	·9887	·99280	·99539	·99926
1·2	·9354	·9628	·9782	·9870	·9922	·99528	·99713	·99971
1·3	·9451	·9700	·9832	·9905	·9946	·99688	·99819	·99986
1·4	·9531	·9756	·9870	·9930	·9962	·99791	·99885	·99989
1·5	·9598	·9800	·9899	·9948	·9973	·99859	·99926	·99999
1·6	·9653	·9836	·9920	·9961	·9981	·99903	·99951	
1·7	·9699	·9864	·9937	·9970	·9986	·99933	·99968	
1·8	·9737	·9886	·9950	·9977	·9990	·99953	·99978	
1·9	·9770	·9904	·9959	·9983	·9992	·99967	·99985	
2·0	·9797	·9919	·9967	·9986	·9994	·99976	·99990	
2·1	·9821	·9931	·9973	·9989	·9996	·99983	·99993	
2·2	·9841	·9941	·9978	·9992	·9997	·99987	·99995	
2·3	·9858	·9950	·9982	·9993	·9998	·99991	·99996	
2·4	·9873	·9957	·9985	·9995	·9998	·99993	·99997	
2·5	·9886	·9963	·9987	·9996	·9998	·99995	·99998	
2·6	·9898	·9967	·9989	·9996	·9999	·99996	·99999	
2·7	·9908	·9972	·9991	·9997	·9999	·99997	·99999	
2·8	·9916	·9975	·9992	·9998	·9999	·99998	·99999	
2·9	·9924	·9978	·9993	·9998	·9999	·99998	·99999	
3·0	·9931	·9981	·9994	·9998	—	·99999	—	—

Voici un zoom sur les premières valeurs :

$z\left(=\frac{x}{s}\right)$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$	$n=9$	$n=10$	For comparison $\left(\frac{\sqrt{7}}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{7x^2}{2}}dx\right)$
·1	·5633	·5745	·5841	·5928	·6006	·60787	·61462	·60411
·2	·6241	·6458	·6634	·6798	·6936	·70705	·71846	·70159
·3	·6804	·7096	·7340	·7549	·7733	·78961	·80423	·78641
·4	·7309	·7657	·7939	·8175	·8376	·85465	·86970	·85520
·5	·7749	·8131	·8428	·8667	·8863	·90251	·91609	·90691
·6	·8125	·8518	·8813	·9040	·9218	·93600	·94732	·94375
·7	·8440	·8830	·9109	·9314	·9468	·95851	·96747	·96799
·8	·8701	·9076	·9332	·9512	·9640	·97328	·98007	·98253
·9	·8915	·9269	·9498	·9652	·9756	·98279	·98780	·99137
1·0	·9092	·9419	·9622	·9751	·9834	·98890	·99252	·99820

Par rapport à notre t moderne, la statistique z de William Gosset est $z = \frac{t}{\sqrt{n-1}}$. Ainsi, pour retrouver la première valeur de la table :

```
pt(0.1 * sqrt(3), df = 3)
```

```
[1] 0.5632413
```

Pour reproduire le début de la table :

```
pz <- function(q, df) {
    round(pt(q = q * sqrt(df - 1), df = df - 1), 4)
}
lignes <- seq(from = 0.1, to = 1, by = 0.1)
colonnes <- 4:10
tablez <- outer(lignes, colonnes, pz)
rownames(tablez) <- as.character(lignes)
colnames(tablez) <- as.character(colonnes)
tablez
```

```
        4      5      6      7      8      9     10
0.1 0.5632 0.5744 0.5840 0.5927 0.6005 0.6078 0.6145
0.2 0.6240 0.6452 0.6633 0.6792 0.6935 0.7064 0.7183
0.3 0.6804 0.7096 0.7340 0.7549 0.7733 0.7896 0.8042
0.4 0.7309 0.7657 0.7940 0.8175 0.8375 0.8547 0.8696
0.5 0.7749 0.8130 0.8428 0.8667 0.8863 0.9025 0.9161
0.6 0.8125 0.8518 0.8813 0.9040 0.9218 0.9359 0.9473
0.7 0.8439 0.8829 0.9109 0.9314 0.9468 0.9585 0.9674
0.8 0.8700 0.9076 0.9332 0.9511 0.9640 0.9733 0.9801
0.9 0.8915 0.9269 0.9498 0.9652 0.9756 0.9828 0.9878
1   0.9092 0.9419 0.9622 0.9751 0.9834 0.9889 0.9925
```

La première tabulation de la loi de Student sous sa forme moderne a été publiée par Fisher dans son livre [2] de 1925 :

TABLE IV.—TABLE OF t

n.	P = ·9.	·8.	·7.	·6.	·5.	·4.	·3.	·2.	·1.	·05.	·02.	·01.
1	·158	·325	·510	·727	1·000	1·376	1·963	3·078	6·314	12·706	31·821	63·657
2	·142	·289	·445	·617	·816	1·061	1·386	1·886	2·920	4·303	6·965	9·925
3	·137	·277	·424	·584	·765	·978	1·250	1·638	2·353	3·182	4·541	5·841
4	·134	·271	·414	·569	·741	·941	1·190	1·533	2·132	2·776	3·747	4·604
5	·132	·267	·408	·559	·727	·920	1·156	1·476	2·015	2·571	3·365	4·032
6	·131	·265	·404	·553	·718	·906	1·134	1·440	1·943	2·447	3·143	3·707
7	·130	·263	·402	·549	·711	·896	1·119	1·415	1·895	2·365	2·998	3·499
8	·130	·262	·399	·546	·706	·889	1·108	1·397	1·860	2·306	2·896	3·355
9	·129	·261	·398	·543	·703	·883	1·100	1·383	1·833	2·262	2·821	3·250
10	·129	·260	·397	·542	·700	·879	1·093	1·372	1·812	2·228	2·764	3·169
11	·129	·260	·396	·540	·697	·876	1·088	1·363	1·796	2·201	2·718	3·106
12	·128	·259	·395	·539	·695	·873	1·083	1·356	1·782	2·179	2·681	3·055
13	·128	·259	·394	·538	·694	·870	1·079	1·350	1·771	2·160	2·650	3·012
14	·128	·258	·393	·537	·692	·868	1·076	1·345	1·761	2·145	2·624	2·977
15	·128	·258	·393	·536	·691	·866	1·074	1·341	1·753	2·131	2·602	2·947
16	·128	·258	·392	·535	·690	·865	1·071	1·337	1·746	2·120	2·583	2·921
17	·128	·257	·392	·534	·689	·863	1·069	1·333	1·740	2·110	2·567	2·898
18	·127	·257	·392	·534	·688	·862	1·067	1·330	1·734	2·101	2·552	2·878
19	·127	·257	·391	·533	·688	·861	1·066	1·328	1·729	2·093	2·539	2·861
20	·127	·257	·391	·533	·687	·860	1·064	1·325	1·725	2·086	2·528	2·845
21	·127	·257	·391	·532	·686	·859	1·063	1·323	1·721	2·080	2·518	2·831
22	·127	·256	·390	·532	·686	·858	1·061	1·321	1·717	2·074	2·508	2·819
23	·127	·256	·390	·532	·685	·858	1·060	1·319	1·714	2·069	2·500	2·807
24	·127	·256	·390	·531	·685	·857	1·059	1·318	1·711	2·064	2·492	2·797
25	·127	·256	·390	·531	·684	·856	1·058	1·316	1·708	2·060	2·485	2·787
26	·127	·256	·390	·531	·684	·856	1·058	1·315	1·706	2·056	2·479	2·779
27	·127	·256	·389	·531	·684	·855	1·057	1·314	1·703	2·052	2·473	2·771
28	·127	·256	·389	·530	·683	·855	1·056	1·313	1·701	2·048	2·467	2·763
29	·127	·256	·389	·530	·683	·854	1·055	1·311	1·699	2·045	2·462	2·756
30	·127	·256	·389	·530	·683	·854	1·055	1·310	1·697	2·042	2·457	2·750
∞	·12566	·25335	·38532	·52440	·67449	·84162	1·03643	1·28155	1·64485	1·95996	2·32634	2·57582

Source : `http://psychclassics.yorku.ca/Fisher/Methods/chap5.htm`

Pour reproduire la table de Fisher :

```
ptbilat <- function(p, df) {
    round(qt(p = 1 - p/2, df = df), 3)
}
```

```
lignes <- c(seq(from = 0.9, to = 0.1, by = -0.1), 0.05, 0.02, 0.01)
colonnes <- 1:30
tablet <- outer(lignes, colonnes, ptbilat)
rownames(tablet) <- as.character(lignes)
colnames(tablet) <- as.character(colonnes)
t(tablet)
```

```
     0.9   0.8   0.7   0.6   0.5   0.4   0.3   0.2   0.1   0.05   0.02   0.01
1  0.158 0.325 0.510 0.727 1.000 1.376 1.963 3.078 6.314 12.706 31.821 63.657
2  0.142 0.289 0.445 0.617 0.816 1.061 1.386 1.886 2.920  4.303  6.965  9.925
3  0.137 0.277 0.424 0.584 0.765 0.978 1.250 1.638 2.353  3.182  4.541  5.841
4  0.134 0.271 0.414 0.569 0.741 0.941 1.190 1.533 2.132  2.776  3.747  4.604
5  0.132 0.267 0.408 0.559 0.727 0.920 1.156 1.476 2.015  2.571  3.365  4.032
6  0.131 0.265 0.404 0.553 0.718 0.906 1.134 1.440 1.943  2.447  3.143  3.707
7  0.130 0.263 0.402 0.549 0.711 0.896 1.119 1.415 1.895  2.365  2.998  3.499
8  0.130 0.262 0.399 0.546 0.706 0.889 1.108 1.397 1.860  2.306  2.896  3.355
9  0.129 0.261 0.398 0.543 0.703 0.883 1.100 1.383 1.833  2.262  2.821  3.250
10 0.129 0.260 0.397 0.542 0.700 0.879 1.093 1.372 1.812  2.228  2.764  3.169
11 0.129 0.260 0.396 0.540 0.697 0.876 1.088 1.363 1.796  2.201  2.718  3.106
12 0.128 0.259 0.395 0.539 0.695 0.873 1.083 1.356 1.782  2.179  2.681  3.055
13 0.128 0.259 0.394 0.538 0.694 0.870 1.079 1.350 1.771  2.160  2.650  3.012
14 0.128 0.258 0.393 0.537 0.692 0.868 1.076 1.345 1.761  2.145  2.624  2.977
15 0.128 0.258 0.393 0.536 0.691 0.866 1.074 1.341 1.753  2.131  2.602  2.947
16 0.128 0.258 0.392 0.535 0.690 0.865 1.071 1.337 1.746  2.120  2.583  2.921
17 0.128 0.257 0.392 0.534 0.689 0.863 1.069 1.333 1.740  2.110  2.567  2.898
18 0.127 0.257 0.392 0.534 0.688 0.862 1.067 1.330 1.734  2.101  2.552  2.878
19 0.127 0.257 0.391 0.533 0.688 0.861 1.066 1.328 1.729  2.093  2.539  2.861
20 0.127 0.257 0.391 0.533 0.687 0.860 1.064 1.325 1.725  2.086  2.528  2.845
21 0.127 0.257 0.391 0.532 0.686 0.859 1.063 1.323 1.721  2.080  2.518  2.831
22 0.127 0.256 0.390 0.532 0.686 0.858 1.061 1.321 1.717  2.074  2.508  2.819
23 0.127 0.256 0.390 0.532 0.685 0.858 1.060 1.319 1.714  2.069  2.500  2.807
24 0.127 0.256 0.390 0.531 0.685 0.857 1.059 1.318 1.711  2.064  2.492  2.797
25 0.127 0.256 0.390 0.531 0.684 0.856 1.058 1.316 1.708  2.060  2.485  2.787
26 0.127 0.256 0.390 0.531 0.684 0.856 1.058 1.315 1.706  2.056  2.479  2.779
27 0.127 0.256 0.389 0.531 0.684 0.855 1.057 1.314 1.703  2.052  2.473  2.771
28 0.127 0.256 0.389 0.530 0.683 0.855 1.056 1.313 1.701  2.048  2.467  2.763
29 0.127 0.256 0.389 0.530 0.683 0.854 1.055 1.311 1.699  2.045  2.462  2.756
30 0.127 0.256 0.389 0.530 0.683 0.854 1.055 1.310 1.697  2.042  2.457  2.750
```

La dernière ligne notée ∞ dans la table correspond à l'approximation par la loi normale :

```
round(qnorm(1 - lignes/2), 5)
```

```
 [1] 0.12566 0.25335 0.38532 0.52440 0.67449 0.84162 1.03643 1.28155 1.64485 1.95996
[11] 2.32635 2.57583
```

On voit que Fisher s'est arrêté à $n = 30$, la fameuse valeur critique pour décréter que l'on a un grand échantillon doit venir de cette table.

On peut trouver sur le web (`http://www.library.adelaide.edu.au/digitised/fisher/stat_tab.pdf`) une reproduction de la table de t publiée dans la sixième édition (1963) de *R. A. Fisher and F. Yates's Statistical Tables for Biological Agricultural and Medical Research (1938)* :

TABLE III. DISTRIBUTION OF t

	Probability.												
n	·9	·8	·7	·6	·5	·4	·3	·2	·1	·05	·02	·01	·001
1	·158	·325	·510	·727	1·000	1·376	1·963	3·078	6·314	12·706	31·821	63·657	636·619
2	·142	·289	·445	·617	·816	1·061	1·386	1·886	2·920	4·303	6·965	9·925	31·598
3	·137	·277	·424	·584	·765	·978	1·250	1·638	2·353	3·182	4·541	5·841	12·924
4	·134	·271	·414	·569	·741	·941	1·190	1·533	2·132	2·776	3·747	4·604	8·610
5	·132	·267	·408	·559	·727	·920	1·156	1·476	2·015	2·571	3·365	4·032	6·869
6	·131	·265	·404	·553	·718	·906	1·134	1·440	1·943	2·447	3·143	3·707	5·959
7	·130	·263	·402	·549	·711	·896	1·119	1·415	1·895	2·365	2·998	3·499	5·408
8	·130	·262	·399	·546	·706	·889	1·108	1·397	1·860	2·306	2·896	3·355	5·041
9	·129	·261	·398	·543	·703	·883	1·100	1·383	1·833	2·262	2·821	3·250	4·781
10	·129	·260	·397	·542	·700	·879	1·093	1·372	1·812	2·228	2·764	3·169	4·587
11	·129	·260	·396	·540	·697	·876	1·088	1·363	1·796	2·201	2·718	3·106	4·437
12	·128	·259	·395	·539	·695	·873	1·083	1·356	1·782	2·179	2·681	3·055	4·318
13	·128	·259	·394	·538	·694	·870	1·079	1·350	1·771	2·160	2·650	3·012	4·221
14	·128	·258	·393	·537	·692	·868	1·076	1·345	1·761	2·145	2·624	2·977	4·140
15	·128	·258	·393	·536	·691	·866	1·074	1·341	1·753	2·131	2·602	2·947	4·073
16	·128	·258	·392	·535	·690	·865	1·071	1·337	1·746	2·120	2·583	2·921	4·015
17	·128	·257	·392	·534	·689	·863	1·069	1·333	1·740	2·110	2·567	2·898	3·965
18	·127	·257	·392	·534	·688	·862	1·067	1·330	1·734	2·101	2·552	2·878	3·922
19	·127	·257	·391	·533	·688	·861	1·066	1·328	1·729	2·093	2·539	2·861	3·883
20	·127	·257	·391	·533	·687	·860	1·064	1·325	1·725	2·086	2·528	2·845	3·850
21	·127	·257	·391	·532	·686	·859	1·063	1·323	1·721	2·080	2·518	2·831	3·819
22	·127	·256	·390	·532	·686	·858	1·061	1·321	1·717	2·074	2·508	2·819	3·792
23	·127	·256	·390	·532	·685	·858	1·060	1·319	1·714	2·069	2·500	2·807	3·767
24	·127	·256	·390	·531	·685	·857	1·059	1·318	1·711	2·064	2·492	2·797	3·745
25	·127	·256	·390	·531	·684	·856	1·058	1·316	1·708	2·060	2·485	2·787	3·725
26	·127	·256	·390	·531	·684	·856	1·058	1·315	1·706	2·056	2·479	2·779	3·707
27	·127	·256	·389	·531	·684	·855	1·057	1·314	1·703	2·052	2·473	2·771	3·690
28	·127	·256	·389	·530	·683	·855	1·056	1·313	1·701	2·048	2·467	2·763	3·674
29	·127	·256	·389	·530	·683	·854	1·055	1·311	1·699	2·045	2·462	2·756	3·659
30	·127	·256	·389	·530	·683	·854	1·055	1·310	1·697	2·042	2·457	2·750	3·646
40	·126	·255	·388	·529	·681	·851	1·050	1·303	1·684	2·021	2·423	2·704	3·551
60	·126	·254	·387	·527	·679	·848	1·046	1·296	1·671	2·000	2·390	2·660	3·460
120	·126	·254	·386	·526	·677	·845	1·041	1·289	1·658	1·980	2·358	2·617	3·373
∞	·126	·253	·385	·524	·674	·842	1·036	1·282	1·645	1·960	2·326	2·576	3·291

L'exercice a été poussé un petit peu plus loin puisque l'on a maintenant la colonne $p = 0.001$ en plus et quelques valeurs intermédiaires supplémentaires pour $n = 40$, $n = 60$ et $n = 120$.

3 Conclusion

L'utilisation de la loi normale comme approximation de la loi de Student est uniquement liée à des contraintes technologiques du XX^{e} siècle :

1. Coût du calcul pour tabuler la loi de Student.
2. Coût de la conception, de l'impression et de la diffusion des tables statistiques.

Ces verrous technologiques ont complètement disparus au XXI^{e} siècle avec l'apparition de logiciels statistiques performants comme R, il n'y a vraiment plus rien qui justifie l'utilisation d'approximations conduisant à des tests non conservatifs, si ce n'est l'inertie de la tradition.

Quand on travaille avec un risque de première espèce $\alpha = 0.05$ l'approximation est bonne à partir de $n = 30$, ce n'est plus du tout le cas avec des risques de première espèce plus faibles.

4 Test empirique du t sur 3000 Criminels

Student nous explique dans son article [7] page 13 :

SECTION VI. *Practical Test of the foregoing Equations.*

Before I had succeeded in solving my problem analytically, I had endeavoured to do so empirically. The material used was a correlation table containing the height and left middle finger measurements of 3000 criminals, from a paper by W. R. Macdonell (*Biometrika*, Vol. I. p. 219). The measurements were written out on 3000 pieces of cardboard, which were then very thoroughly shuffled and drawn at random. As each card was drawn its numbers were written down in a book which thus contains the measurements of 3000 criminals in a random order. Finally each consecutive set of 4 was taken as a sample—750 in all—and the mean, standard deviation, and correlation† of each sample determined. The difference between the mean of each sample and the mean of the population was then divided by the standard deviation of the sample, giving us the z of Section III.

Les données se trouvent en fait à la page 216, et non 219, de l'article de Macdonell [4].

TABLE III. 3000 *Criminals*. Height (feet and inches).

Left Middle Finger (millimetres).	4′ 7″/16—8″/16	8″/16—9″/16	9″/16—10″/16	10″/16—11″/16	5′ 11″/16—0″/16	0″/16—1″/16	1″/16—2″/16	2″/16—3″/16	3″/16—4″/16	4″/16—5″/16	5″/16—6″/16	6″/16—7″/16	7″/16—8″/16	8″/16—9″/16	9″/16—10″/16	10″/16—11″/16	6′ 11″/16—0″/16	0″/16—1″/16	1″/16—2″/16	2″/16—3″/16	3″/16—4″/16	4″/16—5″/16	Totals
9·4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
5	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
7	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
8	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
9	—	—	1	—	1	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3
10·0	1	—	—	1	2	—	2	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	7
1	—	—	—	1	3	1	—	1	1	—	—	—	—	—	—	—	—	—	—	—	—	—	7
2	—	—	2	2	2	1	—	2	—	1	—	—	—	—	—	—	—	—	—	—	—	—	10
3	—	1	1	3	2	2	3	5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	17
4	—	—	1	1	2	3	3	4	3	3	—	—	—	1	—	—	—	—	—	—	—	—	20
5	—	—	—	1	3	7	6	4	3	1	3	1	—	1	—	—	—	—	—	—	—	—	30
6	—	—	—	1	4	5	9	14	6	3	1	—	—	1	—	—	—	—	—	—	—	—	44
7	—	—	1	2	4	9	14	16	15	7	3	1	2	—	—	—	—	—	—	—	—	—	74
8	—	—	—	2	5	6	14	27	10	7	1	2	1	—	—	—	—	—	—	—	—	—	75
9	—	—	—	—	2	6	14	24	27	14	10	4	1	—	—	—	—	—	—	—	—	—	102
11·0	—	—	—	2	6	12	15	31	37	27	17	10	6	—	—	—	—	—	—	—	—	—	163
1	—	—	—	3	3	12	22	26	24	26	24	7	4	1	—	—	—	—	—	—	—	—	152
2	—	—	—	3	2	7	21	30	38	29	27	20	4	1	—	—	—	—	—	—	—	1	183
3	—	—	—	1	—	5	10	24	26	39	26	24	7	2	—	—	—	—	—	—	—	—	164
4	—	—	—	—	3	4	9	29	56	58	26	22	10	11	—	—	—	—	—	—	—	—	228
5	—	—	—	—	—	5	11	17	33	57	38	34	25	11	2	—	—	—	—	—	—	—	233
6	—	—	—	—	2	1	4	13	37	39	48	38	27	12	2	2	—	1	—	—	—	—	226
7	—	—	—	—	—	2	9	17	30	37	48	45	24	9	9	2	—	—	—	—	—	—	232
8	—	—	—	—	1	—	2	11	15	35	41	34	29	10	5	1	—	—	—	—	—	—	184
9	—	—	—	—	1	1	2	12	10	27	32	35	19	10	9	3	1	—	—	—	—	—	162
12·0	—	—	—	—	—	—	1	4	8	19	42	39	22	16	8	2	2	—	—	—	—	—	163
1	—	—	—	—	—	—	—	2	4	13	22	28	15	27	10	4	1	—	—	—	—	—	126
2	—	—	—	—	—	—	1	2	5	6	23	17	16	11	8	1	1	—	—	—	—	—	91
3	—	—	—	—	—	—	—	—	4	8	10	13	20	23	6	5	—	—	—	—	—	—	89
4	—	—	—	—	—	—	1	1	1	2	7	12	4	7	7	1	—	—	1	—	—	—	44
5	—	—	—	—	—	—	—	1	—	1	3	12	11	8	6	8	—	2	—	—	—	—	52
6	—	—	—	—	—	—	—	—	—	1	—	3	5	7	8	6	3	1	1	—	—	—	35
7	—	—	—	—	—	—	—	—	—	1	1	7	5	5	8	2	2	—	—	—	—	—	31
8	—	—	—	—	—	—	—	—	—	—	1	2	3	1	8	5	3	1	1	—	—	—	25
9	—	—	—	—	—	—	—	—	—	—	—	1	2	2	—	1	1	—	—	—	—	—	7
13·0	—	—	—	—	—	—	—	—	—	—	3	—	1	—	1	—	2	1	—	—	—	—	8
1	—	—	—	—	—	—	—	—	—	—	—	1	1	—	—	—	—	—	—	—	—	—	2
2	—	—	—	—	—	—	—	—	—	—	1	1	—	1	—	3	—	—	—	—	—	—	6
3	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2
4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	1
Totals	1	1	6	23	48	90	175	317	393	462	458	413	264	177	97	46	17	7	4	0	0	1	3000
Means	100	103	102·8	107·0	107·8	109·4	110·6	111·8	113·3	114·8	116·5	117·7	118·6	120·1	122·2	123·9	125·9	126·4	127·7	—	—	112	

Lire les données :

```
crim <- read.table(file = "http://pbil.univ-lyon1.fr/R/donnees/criminals1902.txt")
taille <- as.numeric(substr(colnames(crim), 2, 8))
majeur <- as.numeric(rownames(crim))
taille
```

```
 [1] 142.24 144.78 147.32 149.86 152.40 154.94 157.48 160.02 162.56 165.10 167.64
[12] 170.18 172.72 175.26 177.80 180.34 182.88 185.42 187.96 190.50 193.04 195.58
```

```
majeur
```

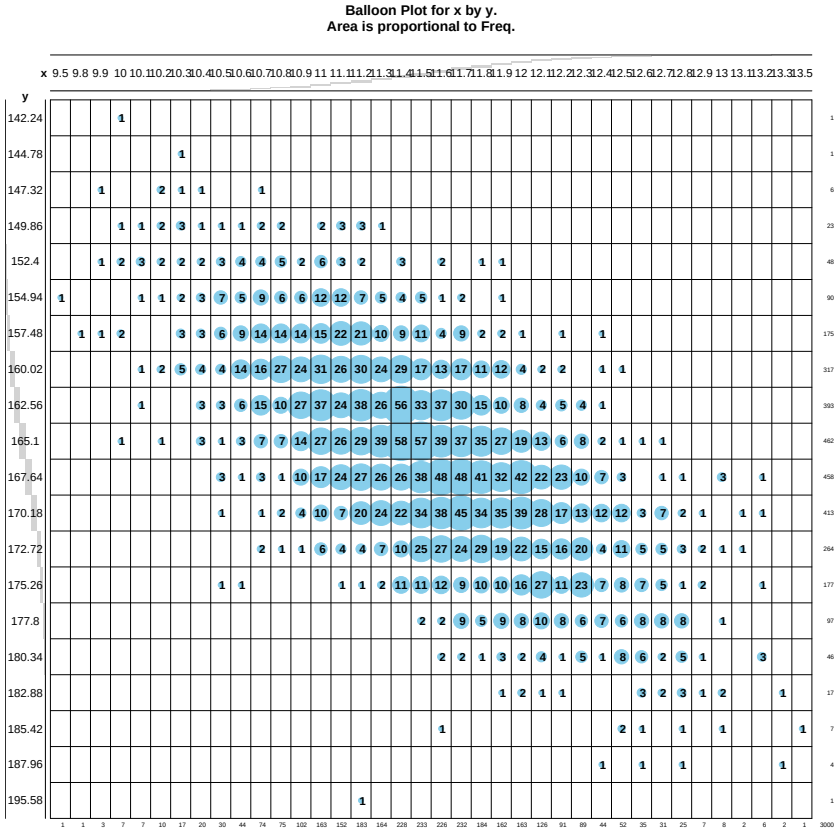
```
 [1]  9.4  9.5  9.6  9.7  9.8  9.9 10.0 10.1 10.2 10.3 10.4 10.5 10.6 10.7 10.8 10.9
[17] 11.0 11.1 11.2 11.3 11.4 11.5 11.6 11.7 11.8 11.9 12.0 12.1 12.2 12.3 12.4 12.5
[33] 12.6 12.7 12.8 12.9 13.0 13.1 13.2 13.3 13.4 13.5
```

Reconstituer les données originales :

```
urne <- data.frame(matrix(NA, nrow = 3000, ncol = 2))
names(urne) <- c("maj", "tai")
n <- 1
for (i in 1:nrow(crim)) {
    for (j in 1:ncol(crim)) {
        if (crim[i, j] != 0) {
            urne[n:(n + crim[i, j] - 1), "maj"] <- majeur[i]
            urne[n:(n + crim[i, j] - 1), "tai"] <- taille[j]
            n <- n + crim[i, j]
        }
    }
}
```

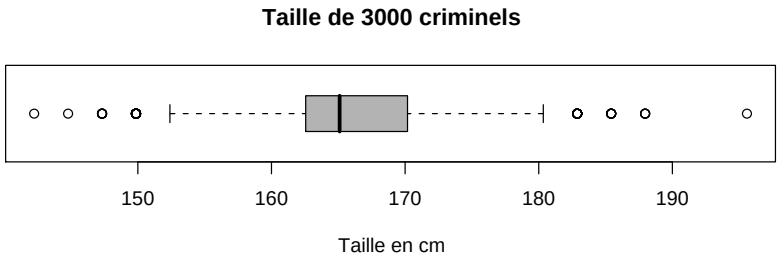
Vérifier que tout va bien :

```
library(gdata)
library(gtools)
library(gplots)
par(cex = 0.75)
balloonplot(table(urne), dotsize = 10)
```



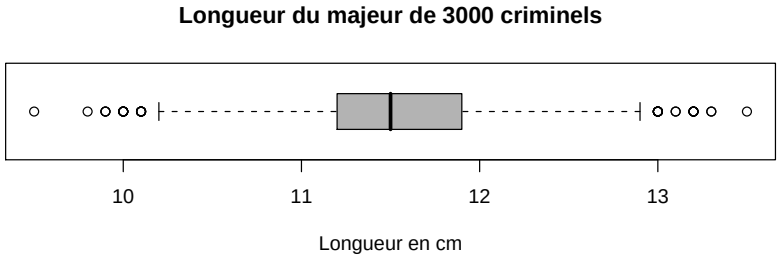
Représenter la taille des individus :

```
boxplot(urne$tai, horizontal = T, col = grey(0.7), xlab = "Taille en cm",
    main = "Taille de 3000 criminels")
```



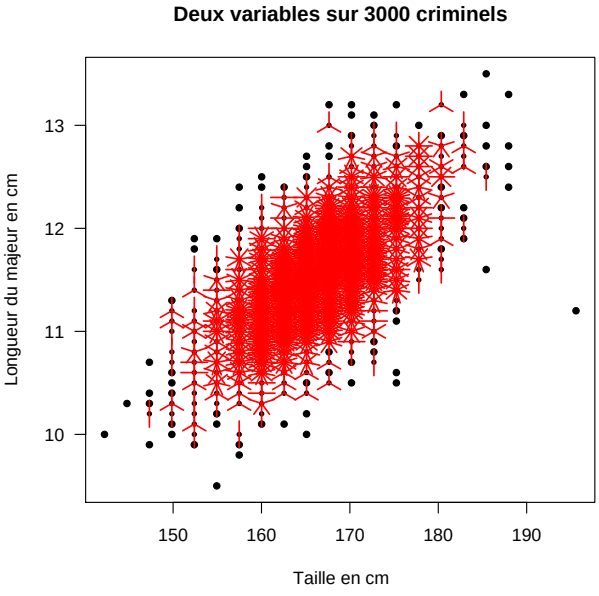
Représenter la longueur des majeurs des individus :

```
boxplot(urne$maj, horizontal = T, col = grey(0.7), xlab = "Longueur en cm",
    main = "Longueur du majeur de 3000 criminels")
```



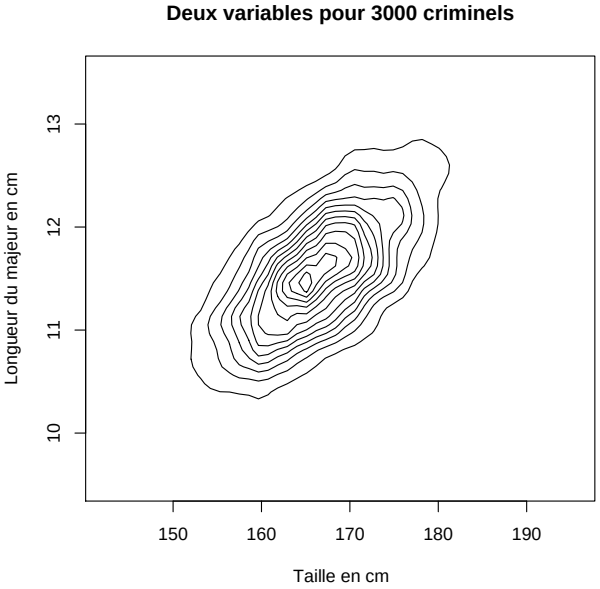
Représenter la corrélation entre les deux variables :

```
sunflowerplot(urne$tai, urne$maj, las = 1, xlab = "Taille en cm",
    ylab = "Longueur du majeur en cm", main = "Deux variables sur 3000 criminels")
```



Faire une représentation graphique en utilisant un estimateur de la densité locale :

```
library(MASS)
contour(kde2d(urne$tai, urne$maj, n = 50), drawlabels = F, xlab = "Taille en cm",
    ylab = "Longueur du majeur en cm", main = "Deux variables pour 3000 criminels")
```



Mélanger les données :

```
mel <- urne[sample(3000), ]
```

Définir un facteur pour chaque échantillon :

```
mel$grp <- as.factor(rep(1:750, each = 4))
head(mel)
```

```
      maj    tai grp
940  11.3 162.56   1
2645 12.2 170.18   1
1616 11.6 167.64   1
438  11.0 160.02   1
951  11.3 162.56   2
2899 12.6 175.26   2
```

Calculer les moyennes des 750 échantillons et représenter graphiquement le résultat :

```
moyennes <- by(mel$maj, mel$grp, mean)
```

Références

[1] R.C. Archibald. A rare pamphlet of Moivre and some of his discoveries. *Isis*, 8 :671–683, 1926.

[2] R.A. Fisher. *Statistical methods for research workers*. Oliver & Boyd, London, U.K., 1925.

[3] Pearson K. Historical note on the origin of the normal curve of errors. *Biometrika*, 16 :402–404, 1924.

[4] W.R. Macdonell. On criminal anthropometry and the identification of criminals. *Biometrika*, 1 :177–227, 1901.

[5] E. Nelson. *Radically elementary probability theory*. Princeton University Press, Princeton, New Hersey, USA, 1987.

[6] G. Pólya. Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem. *Mathematische Zeitschrift*, 8 :171–181, 1920.

[7] Student. The probable error of a mean. *Biometrika*, 6 :1–25, 1908.