Cours de biostatistique ⋈ Illustrations dans 

Décisions et Risques d'erreur

A.B. Dufour & D. Chessel

29 février 2008

La fiche montre la pratique, sur un même jeu de données, de plusieurs ensembles de tests. Les uns sont significatifs, les autres ne le sont pas. Elle permet de se familiariser avec la notion d'hypothèse alternative.

Table des matières

1 Une série de catastrophes

	Date	Type	Lieu	VICT	JOUR	ATT
DEB	01/01/72				1	
1	07/01/72	SA Caravelle	Près d'Ibiza, îles Baléares (Espagne)	104	7	
2	17/01/72	DC 3	A l'atterrissage, Caquenia (Colombie)	1	17	10
3	21/01/72	VC Viscount	Près de Bogota (Colombie)	20	21	4
4	21/01/72	DC 3	Près de San Nicolas (Colombie)	39	21	0
5	26/01/72	DC 9	Explosion d'une bombe, Hermsdorf (RDA)	27	26	5
6	03/02/72	DC 6	Près de Tegal, Java (Indonésie)	6	34	8
7	05/02/72	FH 227	Près de Valledupar (Colombie)	19	36	2
8	11/02/72	DC 4	Disparu (Laos)	23	42	6
9	03/03/72	FH 227	S'écrase sur une maison, près de New-York (USA)	16	63	21
10	14/03/72	SA Caravelle	S'écrase en montagne, près de Al-Fujairah (EAU)	112	74	11
11	19/03/72	DC 9	S'écrase en montagne, près de Aden (Sud-Yemen)	30	79	5
12	13/04/72	N YS 11	S'écrase près de Rio de Janeiro (Brésil)	25	104	25
13	16/04/72	F 27	S'écrase près de Frosinone (Italie)	18	107	3
14	18/04/72	VC 10	Au décollage à Addis-Abeba (Ethiopie)	43	109	2
15	20/04/72	C 46	S'écrase en montagne dans le nord du Pérou	6	111	2
16	05/05/72	DC 8	S'écrase près de Palerme, Sicile (Italie)	115	126	15
17	08/05/72	DC 3	S'écrase au Venezuela	7	129	3
18	18/05/72	AN 10	S'écrase près de Kharkov, Ukraine (URSS)	108	139	10
19	21/05/72	F 27	S'écrase en mer, près de Lobito (Angola)	22	142	3
20	29/05/72	L Constel.	S'écrase près de Cruzairo do Sul (Brésil)	9	150	8
21	14/06/72	DC 8	S'écrase près de Delhi (Inde)	82	166	16
22	15/06/72	CV 880	Explosion d'une bombe au Sud Vietnam	81	167	1
23	18/06/72	HS Trident	S'écrase dans le Surrey (GB)	118	170	3
24	24/06/72	DH Heron	S'écrase près de Ponce (Puerto Rico)	20	176	6
25	29/06/72	DH(C) 6	Collision en vol, près de Appleton (USA)	8	181	5
26	29/06/72	CV 580	Collision en vol, près de Appleton (USA)	5	181	0
27	29/06/72	HFB 320	Au décollage, Blackpool (GB)	7	181	0
28	29/07/72	DC 3	Collision en vol, près de Los Palamos (Colombie)	21	211	30
29	29/07/72	DC 3	Collision en vol, près de Los Palamos (Colombie)	17	211	0
30	11/08/72	F 27	A l'atterrissage à Delhi (Inde)	18	224	13
31	14/08/72	IL 62	S'écrase près de Berlin Est (RDA)	156	227	3
32	16/08/72	DC 3	S'écrase en mer, près de Sandoway (Birmanie)	28	229	2
33	27/08/72	DC 3	S'écrase près de Canaima (Venezuela)	34	240	11
34	01/09/72	Short Skywan	S'écrase sur le mont Giluwe (Nle Guinée)	4	245	5
35	10/09/72	DC 3	S'écrase près de Gondar (Ethiopie)	11	254	9
36	13/09/72	DC 3	S'écrase près de Katmandou (Népal)	31	257	3
37	24/09/72	DC 4	S'écrase près de Saigon (Sud Vietnam)	10	268	11
38	02/10/72	DC 3	Abattu par un mortier à Kampot (Cambodge)	9	276	8
39	02/10/72	IL 18	S'écrase près de Sochi (URSS)	100	276	0
40	13/10/72	IL 62	S'écrase près de Moscou (URSS)	174	287	11
41	21/10/72	N YS 11	S'écrase près d'Athènes (Grèce)	37	295	8
42	27/10/72	VC Viscount	S'écrase près de Clermont-Ferrand (France)	60	301	6
43	30/10/72	F 27	S'écrase près de Poggiorsini (Italie)	27	304	3
44	04/11/72	IL 14	S'écrase près de Polvdiv (Bulgarie)	35	309	5
45	18/11/72	DC 3	Disparu entre l'Islande et Terre-Neuve (Canada)	3	323	14
46	28/11/72	DC 8	Au décollage, à Moscou (URSS)	62	333	10
47	03/12/72	CV 990	Au décollage, à Tenerife, îles Canaries (Espagne)	155	338	5
48	08/12/72	F 27	S'écrase près de Jalkot (Pakistan)	26	343	5
49	08/12/72	B 737	A l'atterrissage à Chicago (USA)	43	343	0
50	20/12/72	DC 9	Collision au sol, Chicago (USA)	10	355	12
51	23/12/72	F 28	S'écrase près d'Oslo (Norvège)	40	358	3
52	29/12/72	L I00I Trister	S'écrase près de Everglades Swamp (USA)	101	364	6
53	22/01/73	B 707	A l'atterrissage, Kano (Niger)	176	388	24
54	29/01/73	IL 18	S'écrase près de Nicosie (Chypre)	37	395	7
55	19/02/73	TU 154	S'écrase près de Pragues (Tchéchoslovaquie)	66	416	21
56	21/02/73	B 727	Abattu dans le désert du Sinai	106	418	2
57	21/02/73	DC 3	S'écrase près de Boquete (Panama)	22	418	0
58	03/03/73	IL 18	S'écrase près de Moscou (URSS)	25	428	10
59	05/03/73	DC 9	S'écrase près de Nantes (France)	68	430	2
60	19/03/73	DC 4	S'écrase près de Ban Me Thuot (Sud Vietnam)	57	444	14
61	10/04/73	VC Vanguard	S'écrase près de Basel (Suisse)	108	466	22
62	05/05/73	B 707	Remous en vol, chaîne des Alpes	1	491	25
63	19/05/73	DC 3	S'écrase près de Svay Rieng (Cambodge)	11	505	14
64	29/05/73	DC 3	S'écrase près de Rimouski (Canada)	4	515	10
65	29/05/73	L Constel.	S'écrase près de Lagoinha (Brésil)	9	515	0
66	31/05/73	B 737	S'écrase près de Delhi (Inde)	48	517	2
67	01/06/73	SA Caravelle	S'écrase près de Sao Luis (Brésil)	23	518	1
68	20/06/73	DC 9	S'écrase près de Puerto Vallarta (Mexico)	27	537	19
69	30/06/73	TU 134	Au décollage, Amman (Jordanie)	2	547	10
70	30/06/73	Con PBY 5A	S'écrase près de Villavicencio (Colombie)	1	547	0

71	11/07/73	B 707	S'écrase près de Paris (France)	123	558	11
72	23/07/73	FH 227	S'écrase près de Saint-Louis (USA)	39	570	12
73	23/07/73	B 707	S'écrase près de Papeete (Tahiti)	79	570	0
74	31/07/73	DC 9	A l'atterrissage, Boston (USA)	89	578	8
75	13/08/73	SA Caravelle	A l'atterrissage, La Coruna (Espagne)	85	591	13
76	22/08/73	DC 3	Disparu en Colombie centrale	16	600	9
77	27/08/73	L Elektra11	S'écrase près de Bogota (Colombie)	42	605	5
78	29/08/73	B 707	Remous vers Los Angeles (USA)	1	607	2
79	08/09/73	DC 8	S'écrase près de King Cove (USA)	6	617	10
80	11/09/73	SA Caravelle	S'écrase près de Titograd (Yougoslavie)	41	620	3
81	28/09/73	CV 600	S'écrase près de Mena (USA)	11	637	17
82	01/10/73	DC 3	S'écrase près de Miriituba (Brésil)	8	640	3
83	13/10/73	TU 104	S'écrase près de Domodedovo (URSS)	28	652	12
84	23/10/73	N YS 11	Au décollage, Rio de Janeiro (Brésil)	5	662	10
85	02/11/73	HP Herald	S'écrase près de Villavicencio (Colombie)	6	672	10
86	03/11/73	DC 10	S'écrase près de Albuquerque (USA)	106	673	1
87	17/11/73	DC 3	S'écrase près de Bato (Sud Vietnam)	27	687	14
88	08/12/73	TU 104	A l'atterrissage, Moscou (URSS)	13	708	21
89	16/12/73	TU 124	S'écrase près de Moscou (URSS)	65	716	8
90	21/12/73	CV 440	S'écrase près de Talara (Pérou)	6	721	5
91	22/12/73	SA Caravelle	S'écrase en montagne, près de Tanger (Maroc)	101	722	1
92	01/01/74	F 28	S'écrase près de Turin (Italie)	38	732	10
93	09/01/74	HS 748	S'écrase près de Florencia (Colombie)	31	740	8
94	10/01/74	DC 4	S'écrase en Bolivie	24	741	1
95	17/01/74	DC 3	S'écrase près de Chigorodo (Colombie)	14	748	7
96	26/01/74	F 28	S'écrase près de Izmir (Turquie)	63	757	9
97	30/01/74	B 707	A l'atterrissage, Pago Pago (Samoas Américaines)	97	761	4
98	22/02/74	V 46	S'écrase près de San Francisco de Moxos (Bolivie)	7	784	23
99	03/03/74	DC 10	Décompression, explosion sur Ermenonville (France)	346	793	9
100	13/03/74	CV 340	S'écrase près de Bishop (USA)	36	803	10
101	15/03/74	SA Caravelle	Au décollage, Téhéran (Iran)	15	805	2
102	04/04/74	DC 4	Au décollage, Fracistown (USA)	78	825	20
103	22/04/74	B 707	S'écrase près de Denfasar, île de Bali (Indonésie)	107	843	18
104	27/04/74	IL 18	S'écrase près de Léningrad (URSS)	118	848	5
105	02/05/74	DC 3	S'écrase dans les Andes (Equateur)	22	853	5
106	06/06/74	VC Viscount	S'écrase près de Cucuta (Colombie)	44	888	35
107	27/06/74	B Stratolin.	Au décollage, Battambang (Cambodge)	19	909	21
108	05/08/74	DC 3	S'écrase près du Mont Apica (Canada)	5	948	39
109	11/08/74	IL 18	S'écrase près de Ouagadougou (Haute Volta)	47	954	6
110	12/08/74	DC 3	S'écrase près de Cali (Colombie)	24	955	1
111	14/08/74	VC Viscount	S'écrase près de Porlamar (Venezuela)	47	957	2
112	07/09/74	F 27	A l'atterrissage, Ptandjungkarang (Indonésie)	33	981	24
113	08/09/74	B 707	Explosion d'une bombe, mer Ionienne	88	982	1
114	11/09/74	DC 9	A l'atterrissage, Charlotte (USA)	70	985	3
115	15/09/74	B 727	Piraterie, près de Phang-Rang (Sud Vietnam)	75	989	4
116	29/10/74	L Elektra 11	S'écrase près de Rae Point (Canada)	32	1033	44
117	20/11/74	B 747	Au décollage, Nairobi (Kenya)	59	1055	22
118	01/12/74	B 727	S'écrase près de Washington DC (USA)	92	1066	11
119	04/12/74	DC 8	S'écrase près de Maskeliya (Sri Lanka)	191	1069	3
120	22/12/74	DC 9	S'écrase près de Maturin (Venezuela)	77	1087	18
121	28/12/74	L 18	Au décollage, Tikal (Guatemala)	24	1093	6
122	29/12/74	AN 24	S'écrase sur les monts Lotru (Roumanie)	32	1094	1
123	16/01/75	AN 2	S'écrase près de San Neua (Laos)	12	1112	18
124	30/01/75	F 28	S'écrase dans la mer de Marmara (Turquie)	42	1126	14
125	03/02/75	HS 748	S'écrase près de Manille (Philippines)	33	1130	4
126	27/02/75	EMB 110	Au décollage, Sao Paulo (Brésil)	15	1154	24
127	12/03/75	DC 4	S'écrase près de Pleiku (Sud Vietnam)	26	1167	13
128	23/04/75	C 46	S'écrase près de Soyari (Bolivie)	4	1209	42
129	24/06/75	B 727	A l'atterrissage, New-York (USA)	113	1271	62
130	31/07/75	VC Viscount	A l'atterrissage, près de Hua Lien (Formose)	27	1308	37
131	31/07/75	YAK 40	S'écrase près de Batumi (URSS)	28	1308	0
132	03/08/75	B 707	S'écrase près d'Agadir (Maroc)	188	1311	3
133	19/08/75	IL 62	A l'atterrissage, Damas (Syrie)	126	1327	16
134	30/08/75	F 27	S'écrase en montagne, île Saint Lawrence (Ales)	10	1338	11
135	01/09/75	TU 134	A l'atterrissage, Leipzig (RDA)	26	1340	2
136	24/09/75	F 28	S'écrase près de Palembang, Sumatra (Indonésie)	26	1363	23
137	27/09/75	CL 44	Au décollage, Miami (USA)	5	1366	3
138	30/09/75	TU 154	S'écrase en mer, près de Beyrouth (Liban)	60	1369	3
139	23/10/75	DH Heron	A l'atterrissage, Cairns (Australie)	11	1392	23
140	30/10/75	DC 9	S'écrase près de Prague (Tchéchoslovaquie)	74	1399	7
141	18/11/75	DC 3	S'écrase à Peten State (Guatemala)	15	1418	19
142	22/11/75	AN 24	S'écrase près de Sofia (Bulgarie)	2	1422	4
FIN	31/12/75				1461	

On va raisonner sur une série d'événements dramatiques[1]. C'est la liste des accidents d'avions survenus entre le 01/01/1972 et le 31/12/1975 (http ://pbil.univ-lyon1.fr/R/donnees/acci.txt). On connaît la date de l'accident, le type de l'appareil impliqué, le lieu et le nombre de victimes (VICT). On a rajouté la date en numéro de jours (JOUR) et le nombre de jours entre deux accidents successifs (ATT).

Le principe de départ est que c'est seulement quand on prend une décision qu'on peut se tromper. En langage traditionnel, cela se dit

Moins tu ouvres la bouche, moins tu as de chances d'avaler une mouche.

Le principe suivant est que, si on prend une décision, on peut se tromper et qu'il vaut mieux éviter de se tromper souvent. En clair, la probabilité de dire une bêtise doit être connue et de préférence petite.

Le reste est une question de technique. Les notions essentielles sont décrites sur des exemples.

2 Combien observe-t-on d'accidents d'avions par période donnée ?

2.1 L'effectif par jour

```
acci <- read.table("http://pbil.univ-lyon1.fr/R/donnees/acci.txt",
    h = T)
jour <- acci$JOUR
```

L'expérience porte sur 1461 jours. Il y a 142 avions accidentés. En moyenne, il y a 0.09719 avions accidentés par jour. On observe :

Avions	0	1	2	3
Jours	1330	121	9	1

Sous l'hypothèse du hasard, on observe une réalisation d'un tirage avec remise de $p = 142$ objets dans $n = 1461$ cases. Cette réalisation est-elle *extraordinaire* ? Est-il normal qu'il y ait des jours avec plusieurs avions concernés ? Le hasard seul permettrait-il la présence de plusieurs accidents le même jour ?

```
echa <- sample(1:1461, 142, replace = T)
```

```
echa
```

```
  [1]  667  129  607  498 1133  360  641  524  585  419   73  200   58 1089  508  433
 [17]  255 1040   57  696   58  365 1118  785  793 1245  953   73   74 1431  375   62
 [33]  679  495  359  805  814  743  433   62 1348  151 1024  573  526  952  972  472
 [49]  498 1269 1254 1118 1246   76  525  276  589 1038 1455  838  779  838  235  672
 [65]  927 1356  760  467    9 1411  919  983 1268  896  506  190 1309  536  805 1452
 [81] 1257  462  712  611  787  110  446  398 1196  130  511 1245 1433 1315  486  888
 [97]  387  754  318  153  796 1411 1181  631 1348  350 1105 1036   18 1413 1114  874
[113] 1371  241  914  576    1 1426  978   45  395  859  711 1294  267 1003  444  705
[129]  816 1022  131 1256  826  661  429  809  970 1446  160 1399 1265  582
```

[1]Source : Eddy P., Potter E. & Page B. (1976). Destination désastre. Grasset, Paris, pp 330-332

```
length(unique(echa))
```

```
[1] 131
```

```
unikacc <- length(unique(echa))
```

Dans un échantillon aléatoire, on a trouvé 131 jours avec accidents, comme dans l'observation. Faisons 1000 expériences de ce type.

```
dis <- rep(0, 1000)
for (i in 1:1000) {
    echa <- sample(1:1461, 142, replace = T)
    dis[i] <- length(unique(echa))
}
table(dis)
```

```
dis
126 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142
  1   3   7  10  24  55  99 143 152 166 141 107  56  30   5   1
```

```
somme <- sum(table(dis)[1:which(names(table(dis)) == "131")])
```

On a trouvé que dans 45 cas sur 1000 (4.5%) le nombre de jours avec accidents était inférieur ou égal au nombre de jours observés. On décide de recommencer l'expérience :

```
dis <- rep(0, 1000)
for (i in 1:1000) {
    echa <- sample(1:1461, 142, replace = T)
    dis[i] <- length(unique(echa))
}
table(dis)
```

```
dis
127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142
  3   2   5  16  31  71  88 127 134 171 150 113  62  21   4   2
```

```
somme <- sum(table(dis)[1:which(names(table(dis)) == "131")])
```

On a trouvé cette fois 57 sur 1000 (5.7%). On recommence encore :

```
dis <- rep(0, 1000)
for (i in 1:1000) {
    echa <- sample(1:1461, 142, replace = T)
    dis[i] <- length(unique(echa))
}
table(dis)
```

```
dis
127 128 129 130 131 132 133 134 135 136 137 138 139 140 141
  1   1   7  14  33  55  97 151 154 163 139 103  62  15   5
```

```
somme <- sum(table(dis)[1:which(names(table(dis)) == "131")])
```

On a trouvé cette fois 56 sur 1000 (5.6%). On recommence encore :

```
dis <- rep(0, 1000)
for (i in 1:1000) {
    echa <- sample(1:1461, 142, replace = T)
    dis[i] <- length(unique(echa))
}
table(dis)
```

```
dis
125 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142
  1   1   4   8  15  20  64 110 114 142 183 153  84  58  35   7   1

 somme <- sum(table(dis)[1:which(names(table(dis)) == "131")])
```

On a trouvé cette fois 49 sur 1000.
Nous avons une indication pour dire que 131 jours avec accidents, c'est plutôt faible mais pas aberrant. Déplaçons la question. Si on tire au hasard, trouve-t-on souvent un jour avec 3 avions accidentés, ou même plus ?

```
 dis <- rep(0, 1000)
 for (i in 1:1000) {
     echa <- sample(1:1461, 142, replace = T)
     dis[i] <- max(table(echa))
 }
 table(dis)

dis
  1   2   3   4
  2 831 164   3

 a1 <- table(dis)[which(names(table(dis)) == "3")]
```

On recommence.

```
dis
  2   3   4
824 170   6
```

On recommence.

```
dis
  1   2   3   4
  1 808 185   6
```

La présence d'au moins 3 catastrophes le même jour n'est pas rare (probabilité autour de 17.3%) et cette observation n'autorise aucune mise en doute du modèle aléatoire. Nous ne savons pas, en regardant les données si le processus d'accidents est aléatoire. Fondamentalement, il y a deux erreurs possibles.
La première est de dire que l'hypothèse est fausse alors qu'elle est vraie. On la trouvera en gros caractères dans le journal le jour où 4 avions s'écrasent dans 4 régions du monde. Le titre énorme sera **"Loi des séries : 4 catastrophes aériennes. Que fait l'organisation mondiale du transport aérien ? "**. Il est possible que ce soit un effet pur du hasard. L'hypothèse est vraie et le journaliste pense que c'est impossible.

> Dire que l'hypothèse est fausse alors qu'elle est vraie, c'est faire une erreur de première espèce.

La seconde est de dire que l'hypothèse est vraie, alors qu'elle est fausse. Le statisticien ne se fera pas prendre en défaut sur la première catégorie car il peut faire le calcul. Il le sera souvent par la seconde. Ici le modèle du hasard est faux car il y a des collisions. Il y a 3 accidents le jour 181 dont 2 forment une collision. Les accidents ne sont pas indépendants (mais très peu, car les collisions sont rares). A retenir :

> Dire que l'hypothèse est vraie alors qu'elle est fausse, c'est faire une erreur de seconde espèce.

Nous allons voir que l'hypothèse est très fausse, mais que pour de nombreux points de vue, elle peut passer pour vraie.

2.2 Le nombre d'accidents par jour - Loi de Poisson

Avions	0	1	2	3
Jours	1330	121	9	1

En mathématiques, on démontre que quand n devient grand (n est le nombre de cases) et que p devient grand (p est le nombre d'objets), le rapport reste constant :

$$\frac{p}{n} = \lambda$$

Avions	0	1	≥ 2
Jours	1330	121	10

Dans notre cas, nous avons $p = 142$ accidents sur $n = 1461$ jours. La valeur de λ est donc : 0.0972.
La probabilité pour qu'une case donnée contienne exactement j objets est :

$$P(j) = e^{-\lambda}\frac{\lambda^j}{j!}$$

Avions	0	1	≥ 2
Probabilités	0.9074	0.0882	0.0044
Effectifs théoriques	1325.6825	128.848	6.4695

L'écart entre les observations (Obs_i) et les prédictions (The_i) se mesure par :

$$D = \sum_i \frac{(Obs_i - The_i)^2}{The_i} = \frac{(1330-1326)^2}{1326} + \frac{(121-129)^2}{129} + \frac{(10-6)^2}{6} = 3.17$$

La probabilité de dépasser cette valeur sous le seul effet du hasard est de 7.5% (loi Khi2 à 1 degré de liberté). Une indication (encore) que les observations ne sont pas tout à fait conformes au modèle aléatoire.

```
1 - pchisq(3.17, 1)
```

```
[1] 0.07500245
```

2.3 L'effectif par an - Loi uniforme discrète

On peut maintenant compter les avions accidentés par année. On trouve 52, 39, 31 et 20. Dans le modèle aléatoire, on a distribué 142 objets dans 4 cases. Le raisonnement ne porte plus sur le nombre d'accidents par jours (1461 observations) mais l'année de l'accident (142 observations). En moyenne, on devrait trouver le même effectif, aux fluctuations aléatoires près, chaque année. Est-ce le cas ?

Année	1972	1973	1974	1975
Observés	52	39	31	20
Probabilités	0.250	0.250	0.250	0.250
Théoriques	35.5	35.5	35.5	35.5

$$D = \sum_i \frac{(Obs_i - The_i)^2}{The_i} = \frac{(52-35.5)^2}{35.5} + \frac{(39-35.5)^2}{35.5} + \frac{(31-35.5)^2}{35.5} + \frac{(20-35.5)^2}{35.5} = 15.35$$

La probabilité d'obtenir un écart au moins aussi grand est de 0.0015 (loi Khi2 à 3 degrés de liberté). Ceci est net. On dit que ce test est très significatif.

On notera alors la remarque : *traiter statistiquement des résultats, ce n'est pas appliquer des tests statistiques, c'est mettre en évidence de l'information.* Le transport aérien voit sa fiabilité augmenter régulièrement pendant la période 1972-1975. Le taux d'accidents tend à décroître significativement. Ce n'est pas la seule technique qui permet de le mettre en évidence.

3 La variable date - Loi uniforme continue

Un accident arbitraire survient à un quelconque moment dans l'intervalle

$$[a = 0,\ b = 1461].$$

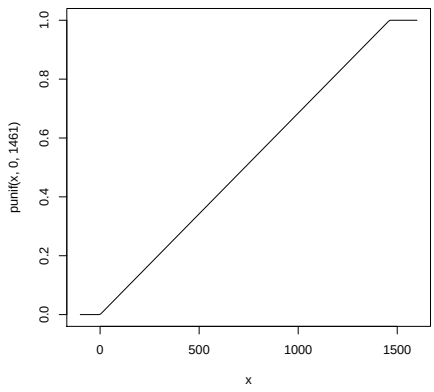
Cela veut dire que la probabilité pour qu'il ait lieu entre la date t_0 et la date t_1 est simplement :

$$P(t_0 < X \leq t_1) = \frac{t_1 - t_0}{b - a}$$

On dit que X suit une loi uniforme sur l'intervalle $[a,b]$. La fonction de répartition est définie par :

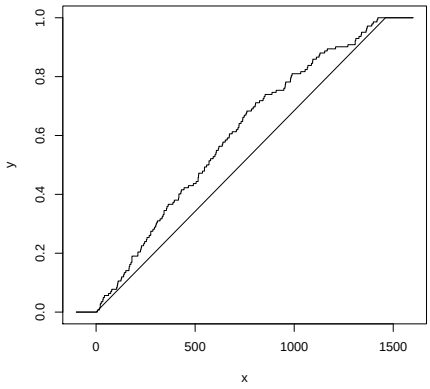
$$\begin{cases} x < a \Rightarrow F(x) = P(X < x) = 0 \\ a \leq x < b \Rightarrow F(x) = P(X < x) = \frac{x-a}{b-a} \\ x \geq b \Rightarrow F(x) = P(X < b) = 1 \end{cases}$$

```
x <- seq(-100, 1600, le = 200)
plot(x, punif(x, 0, 1461), type = "l")
```



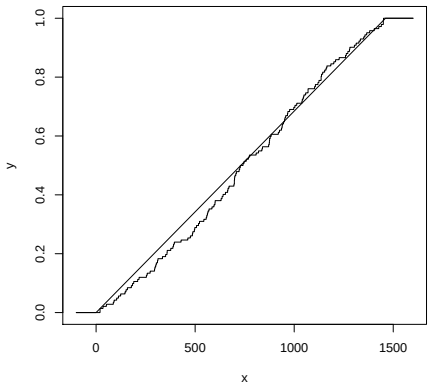
La fonction de répartition empirique est :

```
x <- c(-100, 0, rep(jour[1:142], rep(2, 142)), 1461, 1600)
y <- c(0, 0, 0, (rep(1:141, rep(2, 141))/142), 1, 1, 1)
z <- punif(x, 0, 1461)
plot(x, y, type = "l")
lines(x, z)
```



On a donc un modèle, la fonction de répartition théorique et une observation, la fonction de répartition empirique. L'observation n'est jamais exactement le modèle puisque entre le modèle et la réalisation, il y a une part d'aléatoire. On peut voir une réalisation aléatoire vraie du modèle : il suffit de refaire un tirage.

```
set.seed(1)
echa <- sample(1:1461, 142, replace = T)
echa <- sort(echa)
x <- c(-100, 0, rep(echa[1:142], rep(2, 142)), 1461, 1600)
y <- c(0, 0, 0, (rep(1:141, rep(2, 141))/142), 1, 1, 1)
z <- punif(x, 0, 1461)
plot(x, y, type = "l")
lines(x, z)
```



Si on a trop d'événements au début, la fonction empirique monte trop vite et se trouve au dessus. Si on a trop d'événements à la fin, la fonction théorique ne monte pas assez vite et reste en dessous. Dans le cas où l'hypothèse nulle est vraie, la fonction empirique tourne autour de la théorique et ne s'en éloigne pas trop. Le principe d'un test s'applique à la différence entre l'observation et le modèle. Il peut s'agir d'une différence entre deux valeurs, deux distributions, deux fonctions de répartition, ... La différence existe toujours. La question est de savoir si elle est trop grande pour être acceptable.

3.1 Le test de Kolmogorov-Smirnov

L'écart entre observé et théorique est mesuré en chaque point. Soit $y_i = \frac{x_i - a}{b-a}$ la valeur de la fonction de répartition théorique au point i. Le plus grand écart au dessus de la courbe est mesuré par :

$$D_n^+ = \max_{i=1}^{n} \left\{ \frac{i}{n} - y_i \right\}$$

Le plus grand écart au dessous de la courbe est mesuré par :

$$D_n^- = \max_{i=1}^{n} \left\{ y_i - \frac{i-1}{n} \right\}$$

Le plus grand écart est :

$$D_n = \max \left(D_n^+, D_n^- \right)$$

Les seuils à ne pas dépasser pour un risque α sont tabulés pour les petits échantillons[2] . Pour $n \geq 40$, on a l'approximation :

$$P\left(\sqrt{n} D_n^+ \geq x\right) = e^{-2x^2}$$

```
max0 <- sqrt(142) * max((1:142)/142 - (echa/1461))
max0
```

```
[1] 0.4841519
```

```
exp(-2 * max0 * max0)
```

```
[1] 0.6257489
```

```
neopsim <- exp(-2 * max0 * max0)
```

Pour la simulation, on a 62.57% comme probabilité d'avoir un écart au moins aussi grand. Rien que de très normal.

```
max0 <- sqrt(142) * max((1:142)/142 - (jour/1461))
max0
```

```
[1] 1.933104
```

```
exp(-2 * max0 * max0)
```

```
[1] 0.0005677771
```

```
neopobs <- exp(-2 * max0 * max0)
```

Pour la chronique observée, la probabilité d'avoir un écart au moins aussi grand est de 5 pour 10000. Cela est tout à fait anormal. Le test est très significatif.

[2]Par exemple, G. Saporta (1990) Probabilités, Analyse des Données et Statistique. Editions Technip

3.2 Le test de Cramer-Von-Mises

L'écart entre théorique et observé est mesuré par :

$$T = \frac{1}{12n} + \sum_{i=1}^{n} \left(y_i - \frac{2i-1}{2n} \right)^2$$

Sous l'hypothèse du hasard, pour $n \geq 50$, la probabilité de dépasser 0.4600 vaut 5%, la probabilité de dépasser 0.7378 vaut 1% (*in* G. Saporta, 1990) et la probabilité de dépasser 1.168 vaut 1‰.

```
a0 <- 2 * (1:142) - 1
a0 <- a0/2/142
v0 <- 1/12/142 + sum(((echa/1461) - a0)^2)
v0
```

```
[1] 0.1417855
```

Pour la simulation, la valeur est inférieure à la valeur seuil pour 5%. Elle est non significative.

```
v0 <- 1/12/142 + sum(((jour/1461) - a0)^2)
v0
```

```
[1] 1.361326
```

Pour la chronique observée, la valeur dépasse la valeur seuil pour 1/1000. Elle est très significative. Tous ces résultats sont cohérents.

Remarque. Le test de Cramer prend en compte la somme des carrés des écats obtenus alors que le test de Kolmogorov ne s'appuie que sur la plus grande des différences. Ce dernier est donc plus sensible à l'existence des points aberrants.

3.3 Le test sur la moyenne

Un accident arbitraire intervient dans l'intervalle [a =0, b = 1461] suivant une loi uniforme. La moyenne de cette variable aléatoire est :

$$E(X) = \int_a^b \frac{x-a}{b-a} dx = \left[\frac{(x-a)^2}{2(b-a)} \right]_a^b = \frac{b+a}{2}$$

Remarque : en moyenne il intervient au milieu, rien de bien extraordinaire. La variance vaut :

$$V(X) = \int_a^b \left(\frac{x-a}{b-a} \right)^2 dx - (E(X))^2 = \left[\frac{(x-a)^3}{3(b-a)} \right]_a^b - \left(\frac{b-a}{2} \right)^2 = \frac{(b-a)^2}{12}$$

D'après le théorème central limite, la moyenne des dates d'observations :

$$Y = \frac{1}{n} \sum_{i=1}^{n} X_i$$

suit donc une loi normale de moyenne $E(Y) = \frac{b+a}{2}$ et de variance $V(Y) = \frac{(b-a)^2}{12n}$.

On observe pour la simulation :

⋆ la date moyenne des accidents de la simulation

```
obs <- mean(echa)
obs
```

```
[1] 753.7817
```

⋆ la moyenne théorique

```
m0 <- 1461/2
m0
```

```
[1] 730.5
```

⋆ la variance théorique de la moyenne

```
v0 <- (1461 * 1461)/12/142
sqrt(v0)
```

```
[1] 35.39284
```

⋆ la moyenne observée normalisée

```
z <- (obs - m0)/sqrt(v0)
z
```

```
[1] 0.6578079
```

```
1 - pnorm(abs(z))
```

```
[1] 0.2553308
```

La probabilité de dépasser cette valeur vaut 25.53% et n'a rien d'anormale.

Pour les vraies données :

⋆ la date moyenne des accidents de la simulation

```
obs <- mean(jour)
obs
```

```
[1] 603.7958
```

⋆ la moyenne théorique

```
m0 <- 1461/2
m0
```

```
[1] 730.5
```

⋆ la variance théorique de la moyenne

```
v0 <- (1461 * 1461)/12/142
sqrt(v0)
```

```
[1] 35.39284
```

⋆ la moyenne observée normalisée

```
z <- (obs - m0)/sqrt(v0)
z
```

```
[1] -3.579939
```

```
1 - pnorm(abs(z))
```

```
[1] 0.0001718369
```

Cette valeur est anormale. La probabilité de ne pas dépasser -3.58 vaut 2/10000. Cette valeur est anormalement basse et la moyenne des dates est anormalement faible parce qu'il y a trop d'accidents au début de la chronique. C'est encore une autre façon de voir que la sécurité augmente au cours du temps.
On retiendra que tous ces tests ont en commun une même logique.

1- Formulation de l'hypothèse à tester : " la chronique est un ensemble d'événements indépendants les uns des autres. Les dates sont des variables aléatoires de loi uniforme. "

2- Définition d'une statistique, fonction des observations. Exemples :

$$D = \sum_i \frac{(Obs_i - The_i)^2}{The_i} = 15.35$$

$$\sqrt{n} D_n^+ = \sqrt{n} \max_{i=1}^{n} \left\{ \frac{i}{n} - y_i \right\} = 1.933$$

$$T = \frac{1}{12n} + \sum_{i=1}^{n} \left(y_i - \frac{2i-1}{2n} \right)^2 = 1.361$$

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{b-a}{2}}{\sqrt{\frac{(b-a)^2}{12n}}} = -3.58$$

3- Utilisation de l'information disponible, loi exacte, table numérique, valeurs seuil pour indiquer si la valeur calculée est extraordinairement grande ou petite :

a) 15.35 est la réalisation d'une loi Khi2 à 3 degrés de liberté. Anormalement grand : dépasse la valeur seuil dans la table au seuil de 1%.

b) 1.933 donne une probabilité d'être dépassée de 5/1000, donc est anormalement grande.

c) 1.361 dépasse le seuil donné au risque de 1/1000, donc est anormalement grande.

d) -3.58 est la valeur d'une variable normale réduite, donc anormalement petite (p = 2/10000)

4- Interprétation : transcription en fait du résultat. Ces tests permettent de rejeter l'hypothèse d'un processus uniforme et mettent en évidence l'amélioration continue de la sécurité aérienne (le nombre d'accidents diminue par période, la concentration des accidents diminue de manière continue, la moyenne de la date des accidents est trop faible).

Remarque : La probabilité de rejet (*p-value*) varie fortement d'un test à l'autre. C'est un élément du raisonnement statistique, en rien un résultat expérimental.

4 Le temps d'attente - Loi exponentielle

Nous avons testé le modèle aléatoire par les effectifs par période. Par jour, le processus est peu différent d'un processus dit " Poissonien ". Par an, au contraire, il ne l'est pas. Cela vient de la variation du taux d'accidents dans le temps. Nous avons confirmé largement ce résultat en étudiant la variable date. Nous étudions maintenant le temps d'attente de l'accident suivant.

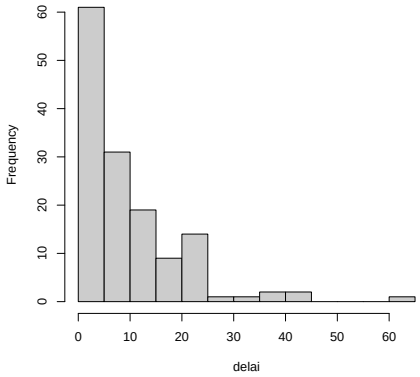
4.1 La distribution du temps d'attente

On appelle processus de renouvellement, une série d'événements qui se produisent avec un délai d'attente du suivant aléatoire mais identique en loi et " sans mémoire ". A chaque accident on recommence à attendre avec la même loi, sans que ce qui s'est passé ne donne de l'indication sur ce qui va se passer. Nous avons 141 observations.

```
diff(jour)
```

```
  [1] 10  4  0  5  8  2  6 21 11  5 25  3  2  2 15  3 10  3  8 16  1  3  6  5  0  0
 [27] 30  0 13  3  2 11  5  9  3 11  8  0 11  8  6  3  5 14 10  5  5  0 12  3  6 24
 [53]  7 21  2  0 10  2 14 22 25 14 10  0  2  1 19 10  0 11 12  0  8 13  9  5  2 10
 [79]  3 17  3 12 10 10  1 14 21  8  5  1 10  8  1  7  9  4 23  9 10  2 20 18  5  5
[105] 35 21 39  6  1  2 24  1  3  4 44 22 11  3 18  6  1 18 14  4 24 13 42 62 37  0
[131]  3 16 11  2 23  3  3 23  7 19  4
```

```
delai <- diff(jour)
hist(delai, col = grey(0.8), main = "")
```

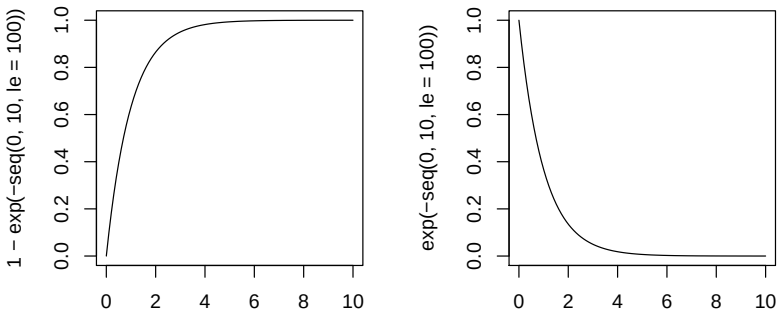


C'est une variable très particulière pour laquelle les petites valeurs sont nombreuses et qui comporte également de grandes valeurs. Une famille de distributions théoriques a ces propriétés. La loi est définie par :

$$P(X < t) = \int_0^t \alpha e^{-\alpha x} dx = 1 - e^{-\alpha t}$$

Pour $\alpha = 1$, la fonction de répartition et la fonction de densité sont :

```
par(mfrow = c(1, 2))
plot(seq(0, 10, le = 100), 1 - exp(-seq(0, 10, le = 100)), type = "l",
    xlab = "")
plot(seq(0, 10, le = 100), exp(-seq(0, 10, le = 100)), type = "l",
    xlab = "")
par(mfrow = c(1, 1))
```



Si X_i et une famille de variables aléatoires indépendantes suivant toutes cette même loi de paramètre α, l'estimation au maximum de vraisemblance de α est donnée simplement par :

$$\alpha = \frac{1}{\frac{X_1+\ldots+X_n}{n}} = \frac{1}{\bar{X}}$$

On trouve :

- le temps d'attente moyen du prochain accident est de 10 jours

```
mean(delai)
```

```
[1] 10.03546
```

```
1/mean(delai)
```

```
[1] 0.09964664
```

- Pour ajuster les observations et le modèle, on définit des classes :

```
bornes <- c(0, 10, 20, 40, 70)
```

- On calcule la valeur de la fonction de répartition pour ces valeurs :

```
alpha <- 1/mean(delai)
alpha
```

```
[1] 0.09964664
```

```
w0 <- 1 - exp(-alpha * bornes)
w0
```

```
[1] 0.0000000 0.6308183 0.8637049 0.9814236 0.9990653
```

- On calcule la probabilité de chaque classe et les effectifs attendus :

```
diff(w0)
```

```
[1] 0.63081833 0.23288656 0.11771875 0.01764164
```

```
w0 <- 141 * diff(w0)
w0
```

```
[1] 88.945385 32.837006 16.598344  2.487471
```

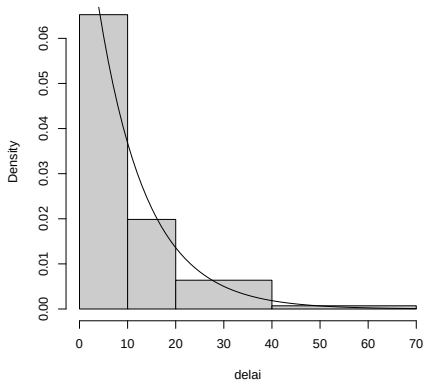
- On compte les effectifs observés par classes :

```
table(cut(delai, bornes, include.lowest = T))
```

```
 [0,10] (10,20] (20,40] (40,70]
     92      28      18       3
```

- On superpose la courbe et l'histogramme en respectant les surfaces :

```
hist(delai, breaks = bornes, proba = T, include.lowest = T, col = grey(0.8),
    main = "")
lines(4:70, dexp(4:70, 1/mean(delai)))
```



⋆ On compare les effectifs théoriques et observés :

Classes	$[0, 10]$	$]10, 20]$	$]20, 40]$	$]40, 70]$
Observés	92	28	18	3
Probabilités	0.631	0.233	0.118	0.018
Théoriques	88.9	32.8	16.6	2.5

⋆ On teste l'écart :

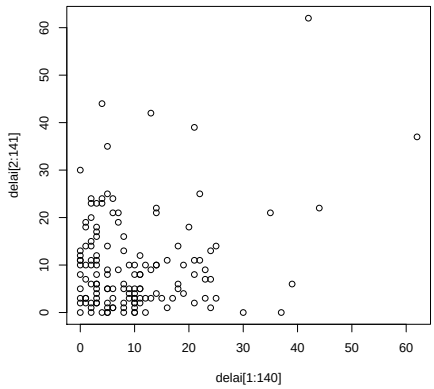
$$D = \sum_i \frac{(Obs_i - The_i)^2}{The_i} = \frac{(92 - 88.9)^2}{88.9} + \frac{(28 - 32.8)^2}{32.8} + \frac{(18 - 16.6)^2}{16.6} + \frac{(3 - 2.5)^2}{2.5} = 1.03$$

La quantité suit une loi Khi2 à 2 degrés de liberté (4 classes -1 contrainte - 1 paramètre estimé). Il ne présente aucun caractère anormal et l'ajustement est bon. Le test est non significatif.

4.2 L'autocorrélation

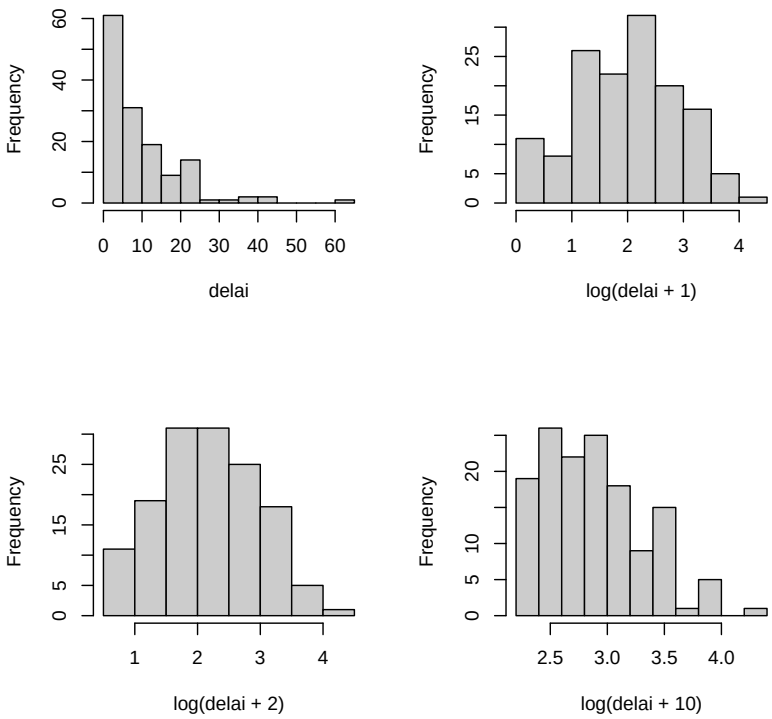
On peut se demander si un temps d'attente important est suivi par un temps plus court ou au contraire si le délai entre deux accidents est indépendant du délai qui précède.

```
plot(delai[1:140], delai[2:141])
```



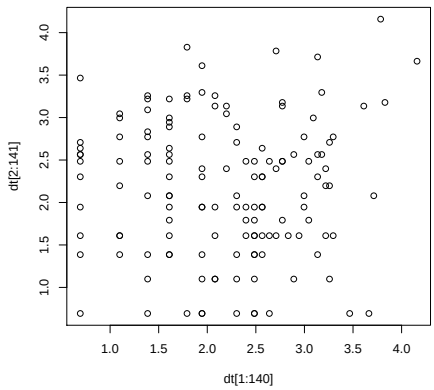
Cette figure n'est pas très explicite car les grandes valeurs déterminent les échelles. On fait un changement de variable.

```
par(mfrow = c(2, 2))
hist(delai, col = grey(0.8), main = "")
hist(log(delai + 1), col = grey(0.8), main = "")
hist(log(delai + 2), col = grey(0.8), main = "")
hist(log(delai + 10), col = grey(0.8), main = "")
par(mfrow = c(1, 1))
```



On choisit la transformation $y = \log(x + 2)$ qui donne à la valeur transformée x (cad le délai) une distribution symétrique.

```
dt <- log(delai + 2)
plot(dt[1:140], dt[2:141])
```



La corrélation entre deux valeurs successives du temps d'attente semble être nulle. Il existe un test basé sur cette notion :

```
cor.test(dt[1:140], dt[2:141])
```

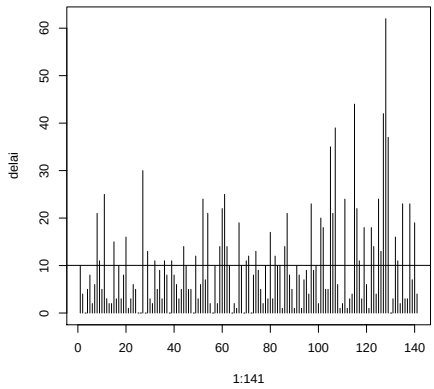
```
        Pearson's product-moment correlation

data:  dt[1:140] and dt[2:141]
t = 0.0746, df = 138, p-value = 0.9406
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1597211  0.1720724
sample estimates:
         cor
0.006350421
```

La mesure de la corrélation vaut 0.00635. Elle a une probabilité d'être dépassée de 0.94. Elle est vraiment sans signification statistique. Alors, les temps d'attente se présentent-ils dans un ordre quelconque ?

4.3 Suites binaires

```
plot(1:141, delai, type = "h")
abline(h = mean(delai))
```



Cette figure représente les temps d'attente dans l'ordre d'observation. Un doute ?

```
sum(delai < 8)
```

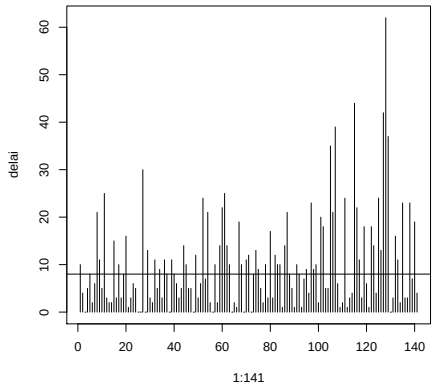
```
[1] 70
```

```
sum(delai >= 8)
```

```
[1] 71
```

70 valeurs sont inférieures à 8 jours, 71 sont supérieures ou égales. 8 est la **médiane**.

```
plot(1:141, delai, type = "h")
abline(h = median(delai))
```



Une valeur sur 2 (71) est au-dessus, une valeur sur 2 (70) est en dessous. Nous sommes dans l'espace des $\binom{141}{70}$ manières de choisir 70 positions (les petits) sur 141 possibles ou, ce qui est la même chose, dans l'espace des $\binom{141}{71}$ manières de choisir 71 positions (les grands) sur 141 possibles. Prenons la première formulation.

```
1 - (delai >= 8)
```

```
  [1] 0 1 1 1 0 1 1 0 0 1 0 1 1 1 0 1 0 1 0 0 1 1 1 1 1 1 0 1 0 1 1 0 1 0 1 0 0 1 0 0
 [41] 1 1 1 0 0 1 1 1 0 1 1 0 1 0 1 1 0 1 0 0 0 0 0 0 1 1 1 0 0 1 0 0 1 0 0 0 1 1 0 1 0
 [81] 1 0 0 0 1 0 0 0 1 1 0 0 1 1 0 1 0 0 0 1 0 0 1 1 0 0 0 1 1 1 0 1 1 1 0 0 0 1 0 1
[121] 1 0 0 1 0 0 0 0 0 1 1 0 0 1 0 1 1 0 1 0 1
```

```
delai < 8
```

```
  [1] FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE
 [14]  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
 [27] FALSE  TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE
 [40] FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE
 [53]  TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE
 [66]  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE
 [79]  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE
 [92] FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE
[105] FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE
[118]  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE
[131]  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
```

Les petits sont les 1 (`TRUE`), les grands sont les 0 (`FALSE`). Il y a $\binom{141}{70}$ séries possibles. Celle-ci est-elle extraordinaire ? Donnons deux points de vue.
Le premier est celui du nombre de suites de caractères identiques :

```
0 1 1 1 1 1 0 0 0 0 1 1 0 0 1 1 1 1 1 1 0 0 0 1 1 0 0
1         2       3   4   5           6     7   8   9
```

Autre manière de compter (+1 à chaque changement +1 à la fin) :

```
0 1 1 1 1 1 0 0 0 0 1 1 0 0 1 1 1 1 1 1 0 0 0 1 1 0 0
  1         2       3   4   5           6     7   8
```

Dans l'espace $\binom{n}{m}$, si $m \geq 10$ et $n - m \geq 10$ le nombre de suites suit approximativement une loi normale de moyenne et variance :

$$E(NS) = \frac{2m(n-m)}{n} + 1 \quad V(NS) = \frac{2m(n-m)(2m(n-m)-n)}{n^2(n-1)}$$

Nous observons 80 suites :

```
nbresui <- 1 - (delai >= 8)
sum(abs(diff(nbresui))) + 1
```

```
[1] 80
```

En moyenne, on en attend :

```
2 * 71 * 70/141 + 1
```

```
[1] 71.49645
```

avec une variance de :

```
2 * 70 * 71 * (2 * 70 * 71 - 141)/141/141/140
```

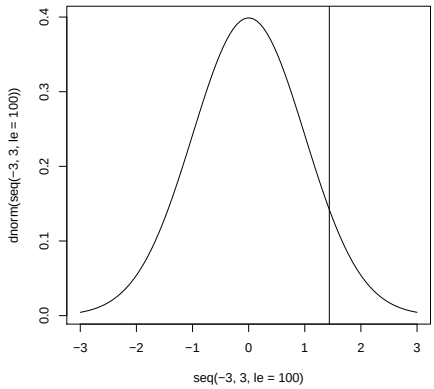
```
[1] 34.99467
```

La valeur normalisée vaut :

```
(80 - 71.5)/sqrt(34.99)
```

```
[1] 1.436968
```

```
plot(seq(-3, 3, le = 100), dnorm(seq(-3, 3, le = 100)), type = "l")
abline(v = 1.437)
```



Cette valeur n'est pas extraordinairement grande ni extraordinairement petite :

$$P(X > 1.437) = 0.075$$
$$P(X < 1.437) = 0.925$$

Elle est dans l'intervalle des valeurs ordinaires $[-1.96, 1.96]$ dans lequel on se trouve dans 95% des cas.

Utilisons un autre point de vue. Les 1 d'une suite binaire ont chacun un numéro d'ordre. La somme de ces numéros est appelée la somme des rangs :

0 0 0 0 1 1 1 1 0 donne 5+6+7+8 = 26

1 0 1 0 1 1 0 0 0 donne 1+3+5+6 = 15

Dans l'espace $\begin{pmatrix} n \\ m \end{pmatrix}$, si $m \geq 10$ et $n - m \geq 10$ la somme des rangs suit approximativement une loi normale de moyenne et variance :

$$E(SR) = \frac{m(n+1)}{2} \quad V(SR) = \frac{m(n-m)(n+1)}{12}$$

Notre suite donne une valeur observée de :

```
sum((1:141) * nbresui)
```

```
[1] 4631
```

En moyenne, on en attend :

```
70 * 142/2
```

```
[1] 4970
```

avec une variance de :

```
70 * 142 * 71/12
```

```
[1] 58811.67
```

La valeur normalisée vaut :

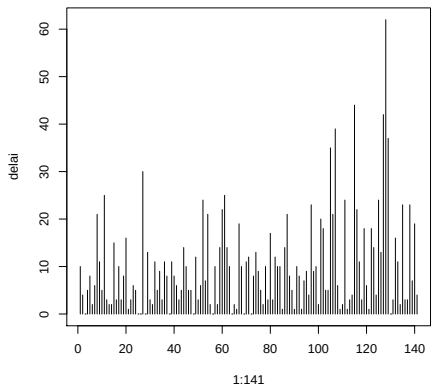
```
(4631 - 4970)/sqrt(58812)
```

```
[1] -1.397870
```

Le test est également non significatif. En réduisant les valeurs à petit/grand, n'avons-nous pas perdu de l'information ?

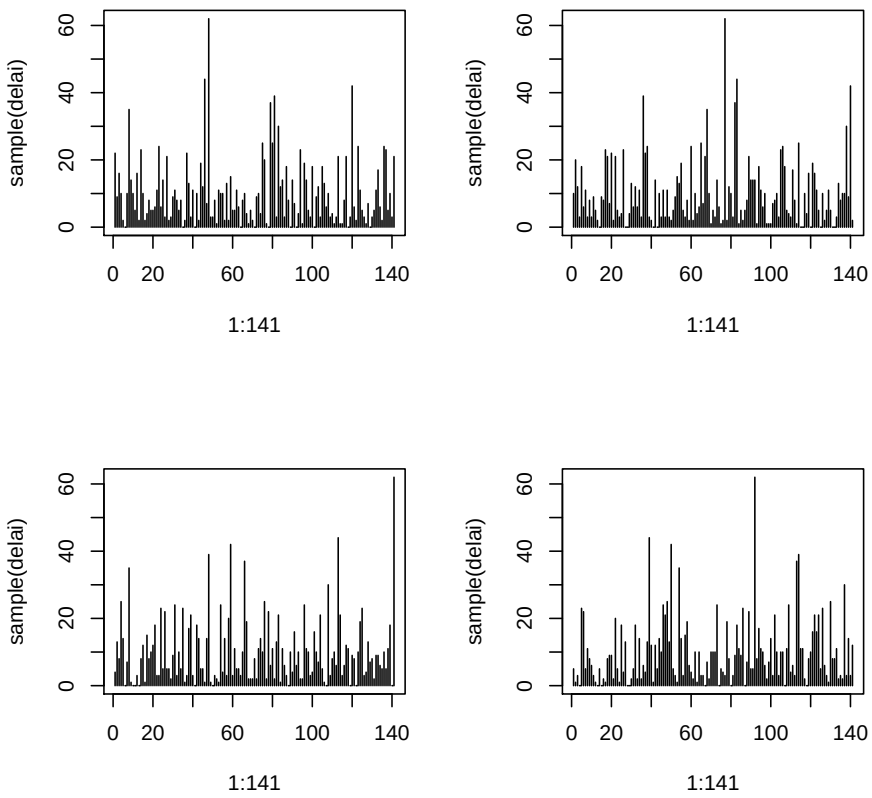
4.4 Espace de permutations

```
plot(1:141, delai, type = "h")
```



Les valeurs observées sont intervenues dans un ordre donné. En voici 4 autres possibles.

```
par(mfrow = c(2, 2))
plot(1:141, sample(delai), type = "h")
plot(1:141, sample(delai), type = "h")
plot(1:141, sample(delai), type = "h")
plot(1:141, sample(delai), type = "h")
```



Nous pourrions en faire une infinité. Il y en a 141 !
Chaque permutation des données permet de calculer la quantité :

$$G = \sum_{i=1}^{n} i x_i$$

Dans l'espace $n!$, si $n \geq 25$ la statistique G suit approximativement une loi normale de moyenne et variance :

$$E(G) = \frac{n(n+1)}{2} m_X \qquad V(G) = \frac{n^2(n+1)}{12} v_X$$

$$m_X = \frac{1}{n} \sum_{i=1}^{n} x_i \quad v_X = \frac{1}{n} \sum_{i=1}^{n} (x_i - m_X)^2$$

Notre suite donne une valeur observée de :

```
sum((1:141) * delai)
```

```
[1] 116185
```

En moyenne, on en attend :

```
141 * 142 * mean(delai)/2
```

```
[1] 100465
```

avec une variance de :

```
141 * 141 * 142 * (var(delai) * 140/141)/12
```

```
[1] 23890956
```

La valeur normalisée vaut :

```
(116185 - 100465)/sqrt(23890956)
```

```
[1] 3.216146
```

```
zs <- (116185 - 100465)/sqrt(23890956)
pnorm(zs)
```

```
[1] 0.9993504
```

```
1 - pnorm(zs)
```

```
[1] 0.0006496232
```

Le test est très significatif. La valeur calculée n'a que 6.5 chances sur 10000 d'être dépassée. Les grands intervalles ont un rang en moyenne trop grand. On retrouve l'indication d'une augmentation de la sécurité.

5 Bilan

Nous pouvons reprendre les résultats qui précèdent.

Hypothèses	**Statistique**	**Résultats**
p objets dans n cases	Nombre de cases occupées	NS
Loi de Poisson	Khi2 d'ajustement	NS
p objets dans n cases	Khi2 d'ajustement	$p = 0.0015$
Distribution uniforme des dates	Kolmogorov	$p = 0.0006$
idem	Cramer-Von-Mises	$p < 0.001$
idem	test sur la moyenne	$p = 0.0002$
Loi exponentielle des délais	Khi2 d'ajustement	NS
Processus de renouvellement	Autocorrélation	NS
m positions sur n	Nombre de suites	NS
m positions sur n	Sommes des rangs	NS
Permutations	Gradient	$p = 0.0006$

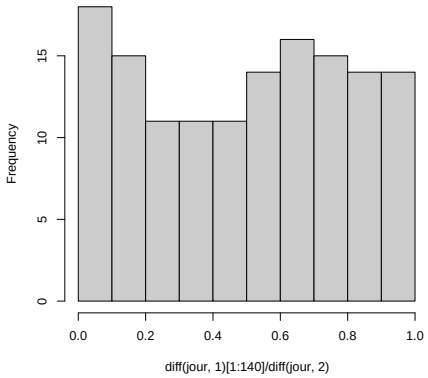
Il est important de savoir que toutes les hypothèses testées sont compatibles : elles dérivent toutes d'un même modèle mathématique qu'on appelle le processus Poissonien. Dans un tel processus qu'on peut dire *aléatoire*, les événements interviennent à tout moment, indépendamment les uns des autres. Si l'intervalle d'observations est divisé en n unités, ils sont répartis au hasard comme p objets dans n cases. Pour un intervalle donné, le nombre d'accidents suit une loi de Poisson et les intervalles successifs forment un échantillon d'une loi de Poisson. Les dates sont indépendantes et ont une distribution uniforme. Ils forment un échantillon d'une loi uniforme. Les temps d'attente entre deux accidents sont indépendants et suivent tous une loi exponentielle. Ils forment un échantillon d'une loi exponentielle. Tous les tests portent donc sur le caractère poissonien du processus.

Une moitié de ces tests sont non significatifs. Une moitié de ces tests sont très significatifs. Y en a t'il des bons et des mauvais ? La contradiction vient du fait qu'il manque encore un élément au raisonnement. Quand on teste une hypothèse, dite hypothèse *nulle*, il faut penser à une *alternative*. On doit tester une hypothèse contre une autre.

La zone de rejet, c'est-à-dire l'événement dont on cherche à savoir s'il est anormalement rare (qui a une petite probabilité de survenir sous l'hypothèse nulle) doit avoir une forte probabilité d'intervenir si l'alternative est vraie. Quand on observe un nombre aléatoire compris entre 1 et 100, on a 1 chance sur 100 d'observer disons 23. Ce n'est pas parce que l'observation est rare qu'on a un test statistique. Si l'alternative est " le nombre est grand ", l'événement {96, 97, 98, 99, 100} formera la zone de rejet à 5%. Si l'alternative est " le nombre est petit ",", l'événement {1, 2, 3, 4, 5} formera la zone de rejet à 5%. Si l'alternative est " le nombre est voisin de 25 " l'événement {23, 24, 25, 26, 27} formera la zone de rejet à 5%. L'alternative doit absolument être formulée avant le test, sinon il y a escroquerie.

Nous avons en fait testé deux alternatives. **La première** est le caractère *localement aléatoire* du processus. Au cour d'une petite période centrée sur l'instant t le nombre d'accidents suit une loi de Poisson de paramètre $\lambda(t)$. Le temps d'attente au suivant suit une loi exponentielle de paramètre $\alpha(t)$. Deux temps d'attente successifs sont indépendants. Deux effectifs successifs sont indépendants. Les ajustements peu sensibles aux variations modérées de $\lambda(t)$ ou $\alpha(t)$ disent OUI. On peut faire un test centré directement sur cet aspect. Si deux temps d'attente consécutifs suivent une loi exponentielle de même paramètre, l'accident du milieu tombe au hasard dans l'intervalle entre le premier et le troisième :

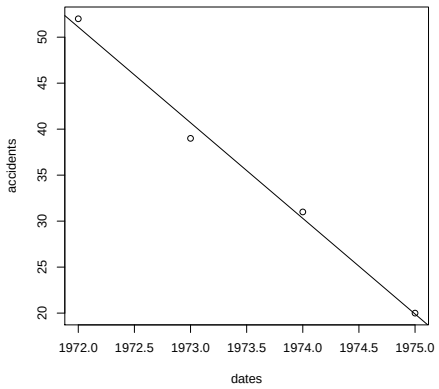
```
hist(diff(jour, 1)[1:140]/diff(jour, 2), col = grey(0.8), main = "")
```



C'est tout à fait acceptable. Tous les tests centrés sur cette alternative sont non significatifs parce que cette hypothèse est vraie. L'alternative " il y a localement des agrégats de catastrophes " comme l'alternative " il y a une certaine régularité des accidents " sont fausses.

La seconde porte sur la densité des catastrophes. L'hypothèse nulle est $\alpha(t) = Cte$ ou $\lambda(t) = Cte$. Tous les tests centrés sur cette hypothèse sont très significatifs. Il y a trop d'accidents la première année, les dates petites sont trop nombreuses (accidents du début), les temps d'attente sont trop grands sur la fin. L'hypothèse nulle est rejetée au profit de l'alternative d'un paramètre non constant. Le nombre d'accidents par an suffit :

```
dates <- c(1972, 1973, 1974, 1975)
accidents <- c(52, 39, 31, 20)
plot(dates, accidents)
abline(lm(accidents ~ dates))
```



On dit que les tests de la première série *sont puissants contre l'alternative d'un caractère localement non poissonien* alors que les tests de la seconde série *sont puissants contre l'alternative d'un caractère globalement non uniforme.*

Le résultat de cette étude est que cette série d'événements est un processus aléatoire à intensité décroissante.

En résumé : faire un test statistique, c'est choisir une hypothèse nulle, une statistique et une zone de rejet peu probable (p) quand l'hypothèse nulle est vraie et probable quand une hypothèse alternative précisée est vraie.

1. Si la valeur calculée tombe dans la zone de rejet, on rejette l'hypothèse nulle au profit de l'alternative. Si l'hypothèse nulle est fausse, tant mieux. Si elle est vraie, on a commis une erreur de première espèce. La probabilité de se tromper est p. Si p est très faible, pas de problème. Si p n'est pas très faible et qu'on risque sa tête, il vaut mieux réfléchir encore.
2. Si la valeur calculée ne tombe pas dans la zone de rejet, on accepte l'hypothèse nulle. Si elle est vraie, tant mieux. Si elle est fausse, on a commis une erreur de seconde espèce. Si on sait calculer son risque, on prend une décision sérieuse. Si on n'a aucune idée du risque de se tromper, il vaut mieux ne rien dire.