R HANDBOOK

Synthèse : LATIF Mehdi
mehdi.latif@etu.univ-nantes.fr

Année 2016/2017

Table des matières

I Université FUN
Introduction à la statistique avec R

Bruno Falissard, Christophe Lalanne

Lien vers le cours de **l'Université FUN** : ici

II Université de Nantes

III Théorie de la statistique

1 Sur l'adéquation à une loi de probabilité avec R - Christophe Chesneau

Auteur : Christophe Chesneau

Note : L'objectif de ce document est de présenter les principaux outils statistiques et commandes **R** utilisés pour juger de l'adéquation de la distribution des valeurs d'un caractère à une loi de probabilité.

1.1 Point de départ

On observe la valeur d'un caractère X pour chacun des n individus d'un échantillon. Ces observations constituent les données : $x_1; \ldots; x_n$. On modélise alors X comme une *var*[1] (en gardant la notation X par convention). Soit $\mathcal{L}$ une loi de probabilité étant possiblement en adéquation avec la loi inconnue de X. La problématique est la suivante :

Est-ce que ces données nous permettent d'affirmer que X ne suit pas la loi $\mathcal{L}$ (avec un faible risque de se tromper) ?

Pour répondre à cette question, on distingue deux approches complémentaires :
- Analyses graphiques.
- Tests statistiques adaptées reposant sur les hypothèses :

$$H_0 : \text{" } X \text{ suit la loi } \mathcal{L} \text{ " } \textbf{contre } H_1 : \text{" } X \text{ ne suit pas la loi } \mathcal{L} \text{ ".}$$

Dans ce document, nous mettons en oeuvre ces tests en utilisant la p-valeur.
La p-valeur est le plus petit réel $\alpha \in]0;1[$ calculé à partir des données tel que l'on puisse se permettre de rejeter H_0 au risque $100\alpha\%$. Autrement écrit, la p-valeur est une estimation ponctuelle de la probabilité critique de se tromper en rejetant H0 alors que H_0 est vraie.

Les logiciels actuels travaillent principalement avec cette p-valeur.

1.2 Analyses graphiques

Cas de caractères qualitatifs ou quantitatifs "discrets"

Soit X un caractère non chiffré (qualitatif) ou chiffré (quantitatif) prenant un ensemble dénombrable de valeurs (possiblement infini). L'analyse graphique la plus pertinente pour juger de l'adéquation de la loi de X avec $\mathcal{L}$ repose sur le schéma suivant :
- On trace le barplot des fréquences correspondantes aux données.
- On superpose les valeurs de la "densité" associée à la loi $\mathcal{L}$ en estimant éventuellement les paramètres inconnus de celle-ci.

1. Variable Aléatoire

Exemple

1. On souhaite savoir si les entrées à l'hôpital pour une certaine maladie sont réparties au hasard dans l'année ou bien si certains mois sont plus propices à la maladie. On examine le mois d'entrée d'un échantillon de 120 porteurs de la maladie étudiée. Les résultats sont :

Mois d'entrée	1	2	3	4	5	6	7	8	9	10	11	12
Nombre d'entrées	18	16	8	10	6	4	4	9	11	10	12	12

Peut-on affirmer que *"les entrées ne se font pas au hasard dans l'année"* (donc que *"certains mois sont plus propices à la maladie"*) ?

Solution : Soit X la *var* égale au mois d'entrée à l'hôpital d'un porteur de la maladie. Par l'énoncé, on observe la valeur de X sur chacun des n individus (porteurs de la maladie) d'un échantillon avec $n = 120$: $(x_1; \dots; x_n)$ (avec $x_i \in \{1; \dots; 12\}$). On forme alors un vecteur des effectifs :

$$(n_1, n_2, \dots, n_{12}) = (18, 16, \dots, 12)$$

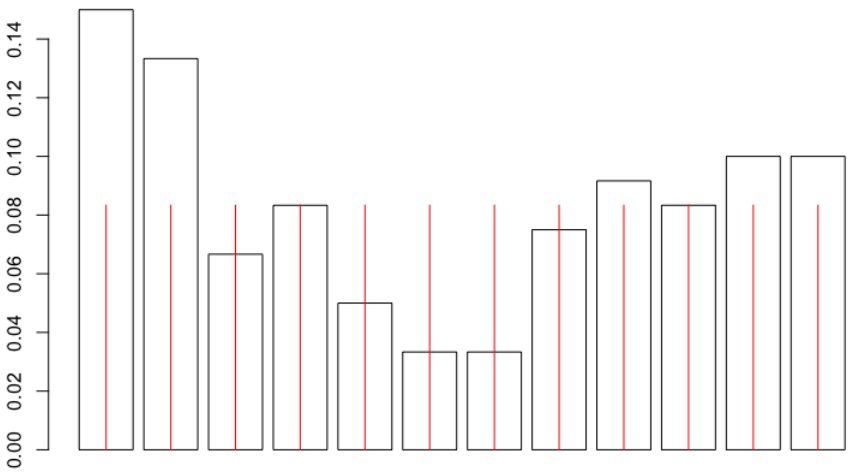
Dire que les entrées se font au hasard dans l'année signifie que X suit la loi uniforme $U(\{1; \dots; 12\})$:

$$\mathbb{P}(X = i) = \frac{1}{12} \text{ avec } i \in \{1, \dots, 12\}$$

La problématique est la suivante : *est-ce que ces données nous permettent d'affirmer que X ne suit pas la loi $\mathcal{L} = \mathcal{U}(\{1; \dots; 12\})$?*
Pour tout $i \in \{1, \dots, 12\}$, une estimation (ponctuelle) de $\mathbb{P}(X = i)$ est la fréquence n_i/n.
Ainsi, pour une analyse graphique, on peut :
— tracer le barplot des fréquences correspondantes aux données
— superposer les valeurs de la "densité" associée à la loi uniforme $\mathcal{U}(\{1; \dots; 12\})$.
On propose les commandes :

```
> nb = c(18, 16, 8, 10, 6, 4, 4, 9, 11, 10, 12, 12)
> bar = barplot(nb / 120, col = "white")
> points(bar, rep(1 / 12, 12), type = "h",col="red")
```

Cela renvoie :



Les différences observées laissent penser que X ne suit pas une loi uniforme.

2. Dans un verger, on étudie le comportement des insectes quand ceux-ci attaquent les fruits.
Soit X le caractère qui dénombre le nombre d'attaques d'insectes sur un fruit pris au hasard. Une étude statistique antérieure montre que, si les attaques se font de façon indépendante les unes des autres, on peut modéliser X comme une *var* suivant une loi de Poisson $\mathcal{P}(\lambda)$ avec λ inconnu.
On considère un échantillon de 300 fruits et on compte le nombre d'attaques sur le fruit. Les résultats sont :

Nombre d'attaques	0	1	2	3	4	5	6	7
Nombre de fruits attaqués	60	105	65	47	15	4	3	1

Peut-on dire que le comportement des insectes est grégaire ? (on dit que le comportement des insectes est grégaire quand chacun d'entre eux a tendance à se comporter comme le voisin et à attaquer le même fruit ; dans ce cas, X ne suit pas une loi de Poisson).

Solution : Par l'énoncé, on observe la valeur de X sur chacun des n individus (fruits) d'un échantillon avec $n = 300$: $(x_1; \dots; x_n)$ (avec $x_i \in \mathbb{N}$). On forme alors un vecteur des effectifs :

$$(n_1; n_2; \dots; n_8) = (60; 105; \dots; 1)$$

Dire que X suit la loi de Poisson $\mathcal{P}(\lambda)$, avec λ inconnu :

$$\mathbb{P}(X = i) = e^{-\lambda}\frac{\lambda^k}{i!},\ i \in \mathbb{N}$$

Pour évaluer la loi de Poisson la plus adaptée à notre contexte, il faut estimer λ à l'aide des données.
Comme $\mathbb{E}(X) = \lambda$, la méthode des moments nous assure qu'une estimation de λ est :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{0\times 60 + 1\times 105 + 2\times 65 + 3\times 47 + 4\times 15 + 5\times 4 + 6\times 3 + 7\times 1}{300} = 1.603333$$

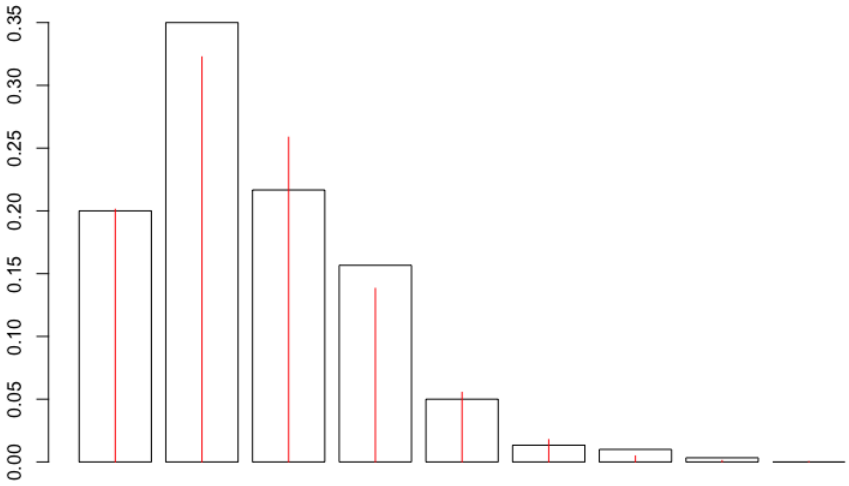
Ainsi, la problématique est la suivante : *est-ce que ces données nous permettent d'affirmer que X ne suit pas la loi* $\mathcal{L} = \mathcal{P}(1.603333)$?
Pour tout $i \in \{0, \dots, 7\}$, une estimation de $\mathbb{P}(X = i)$ est la fréquence n_{i+1}/n et une estimation de $\mathbb{P}(X \geq 8) = 1 - \mathbb{P}(x \leq 7)$ est :

$$1 - \sum_{i=1}^{8} \frac{n_i}{n}$$

Ainsi, pour une analyse graphique, on peut :
— tracer le barplot des fréquences correspondantes aux données
— superposer les valeurs de la "densité" associée à la loi de poisson $\mathcal{P}(1.603333)$.
On propose les commandes :

```
> nb = c(60, 105, 65, 47, 15, 4, 3, 1, 0)
> bar = barplot(nb / 300, col = "white")
> lambda = sum((0:8) * nb) / 300;lambda
[1] 1.603333
> prob = c(dpois(0:7, lambda), 1 - ppois(7, lambda));prob
[1] 0.2012246500 0.3226301888 0.2586418681 0.1382297095 0.0554070752
[6] 0.0177672021 0.0047477912 0.0010874703 0.0002640448
> points(bar, prob, type = "h",col="red")
```

Cela renvoie :



Les différences observées laissent penser que X suit une loi de Poisson, remettant ainsi en cause le comportement grégaire des insectes.

Cas de caractères quantitatifs "continus"

Soit X un caractère chiffré (quantitatif) prenant un ensemble indénombrable de valeurs. Les analyses graphiques possibles pour juger de l'adéquation de la loi de X avec $\mathcal{L}$ sont nombreuses. Les méthodes usuelles sont :

— **Méthode de l'histogramme :** On trace l'histogramme des fréquences correspondantes aux données.
Puis on superpose les valeurs de la "densité" associée à la loi $\mathcal{L}$ en estimant éventuellement les paramètres inconnus de celle-ci.

— **Méthode de la fonction de répartition :** On trace le graphe de la fonction de répartition empirique définie par :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{x_i \leq x\}}, x \in \mathbb{R}$$

Puis on superpose le graphe de la fonction de répartition associée à la loi $\mathcal{L}$ en estimant éventuellement les paramètres inconnus de celle-ci.

— **Méthode de l'approximation de la densité :** On utilise une estimation de la densité inconnue. Puis on superpose les valeurs de la "densité" associée à la loi $\mathcal{L}$ en estimant éventuellement les paramètres inconnus de celle-ci.

— **Méthode du QQ plot (quantile-quantile plot) :** Cette méthode consiste en la comparaison des quantiles empiriques et des quantiles théoriques.
Soit $F(x) = P(X \leq x)$, la fonction de répartition de X et x_p le quantile d'ordre p définie par :

$$x_p = \inf\{x \in \mathbb{R}; F(x) \geq p\}$$

Soient $x_{(1)}; x_{(2)}; \ldots; x_{(n)}$ les données rangées par ordre croissant : $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$. Alors on

peut écrire la fonction de répartition comme :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{x_i \leq x\}} = \begin{cases} 0 & \text{si} \quad x < x_{(1)} \\ \frac{k}{n} & \text{si} \quad x_{(k)} \leq x < x_{(k+1)}, k \in \{1, \dots, n-1\} \\ 1 & \text{si} \quad x \geq x_{(n)} \end{cases}$$

Soit $(p_1; p_2; \dots; p_n)$ une suite strictement croissante de n réels vérifiant $p_k \in](k-1)/n; k/n[$, $k \in \{1; \dots; n\}$, de sorte que $\inf\{x \in \mathbb{R}; F_n(x) \geq p_k\} = x_{(n)}$ pour tout $k \in \{1; \dots; n\}$. On appelle QQ plot le nuage de points $\mathcal{N}$ dans le repère orthonormé $(O; I; J)$ défini par :

$$\mathcal{N} = \{(x_{p_1}, x_{(1)}), (x_{p_2}, x_{(2)}), \dots, (x_{p_n}, x_{(n)})\}$$

Si X suit la loi $\mathcal{L}$, les données font que F_n est une bonne estimation de F et, a fortiori, $x_{(k)}$ doit bien estimer x_{p_k} : $x_{(k)} \simeq x_{p_k}$ pour tout $k \in \{1; \dots; n\}$; les points du nuage $\mathcal{N}$ doivent être proche de la "droite diagonale" d'équation : $y = x$.

— **Cas d'une loi normale : méthode du QQ plot (QQ norm) avec droite de Henry :** Soit z_p le quantile d'ordre p d'une var Z suivant la loi normale centrée réduite $\mathcal{N}(0; 1)$.
Alors, si X suit la loi normale $\mathcal{N}(\mu, \sigma^2)$, le quantile d'ordre p de X vérifie :

$$x_p = \mu + \sigma z_p$$

Par conséquent, au lieu du QQ plot standard, on peut se contenter de construire le nuage de points $\mathcal{N}_*$ dans le repère orthonormé $(O; I; J)$ défini par :

$$\mathcal{N}_* = \{(z_{p_1}, x_{(1)}), (z_{p_2}, x_{(2)}), \dots, (z_{p_n}, x_{(n)})\}$$

Si X suit la loi $\mathcal{N}(\mu, \sigma^2)$, alors les points du nuage $\mathcal{N}_*$ doivent être proche de la droite d'équation :

$$y = \bar{x} + sx$$

avec :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ et } s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Cette droite est appelée droite de Henry.

— **Cas d'une loi normale : méthode de la boîte à moustaches :** On fait la boîte à moustache associée aux données. Si X suit une loi normale, il n'y a approximativement que $0,7\%$ des points qui se trouvent en dehors des "moustaches". D'autre part, la boîte doit être à peu près symétrique par rapport à la médiane, idem pour les moustaches.

Exemples :

1. On fait passer à 50 adolescents le test psychologique de Rorschach. Les temps de passation en minutes du test sont :

43	48	65	55	51	51	44	51	59	62
45	53	55	55	49	34	52	69	45	54
59	36	36	29	52	59	41	58	54	55
72	53	52	49	57	42	70	58	42	53
57	68	40	65	54	49	32	56	50	59

On s'interroge pour savoir si la *var* X qui à un adolescent associe son temps de passation au test suit ou non une loi normale.

Solution : Par l'énoncé, on observe la valeur de X sur chacun des n individus (adolescents) d'un échantillon avec $n = 50$: $(x_1; \dots; x_n)$ (avec $x_i \in \mathbb{R}$).
Dire que X suit une loi normale $\mathcal{N}(\mu, \sigma^2)$, avec μ et σ inconnus, signifie qu'elle possède la densité :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \ x \in \mathbb{R}$$

Pour préciser la loi normale $\mathcal{N}(\mu, \sigma^2)$ la plus adaptée à notre contexte, il faut estimer μ et σ à l'aide des données. On estime alors μ par

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = 51.94$$

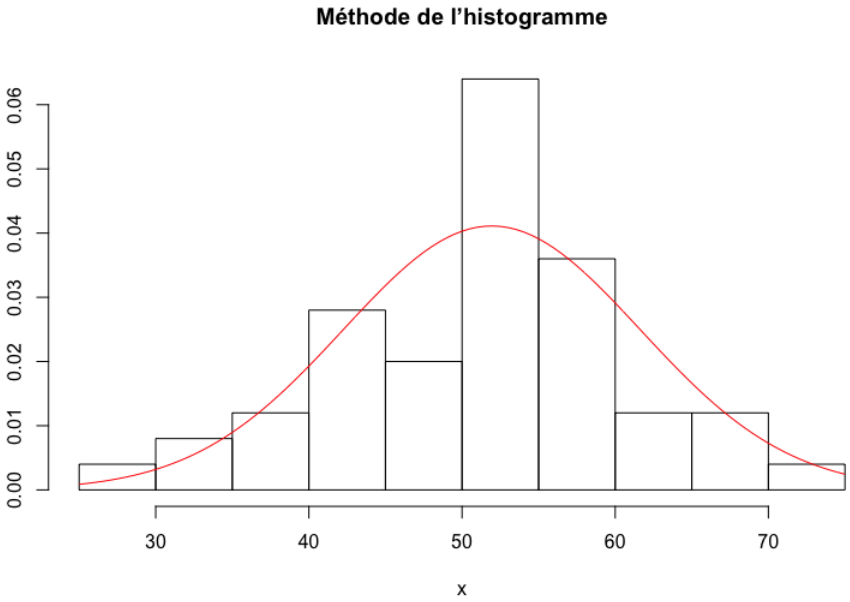
et σ par

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} = 9.704638$$

Ainsi, la problématique est la suivante : *est-ce que ces données nous permettent d'affirmer que X ne suit pas la loi $\mathcal{L} = \mathcal{N}(51,94; 9,704638^2)$?*
Nous allons faire une analyse graphique en utilisant les méthodes présentées précédemment.
— **Méthode de l'histogramme :** On propose les commandes :

```
> x = c(43, 48, 65, 55, 51, 51, 44, 51, 59, 62, 45, 53, 55, 55, 49, 34, 52,
+       69, 45, 54, 59, 36, 36, 29, 52, 59, 41, 58, 54, 55, 72, 53, 52, 49, 57, 42,
+       70, 58, 42, 53, 57, 68, 40, 65, 54, 49, 32, 56, 50, 59)
> hist(x, freq = FALSE, main = "Méthode de l'histogramme", ylab = "")
> curve(dnorm(x, 51.94, 9.704638), add = TRUE, col="red")
```

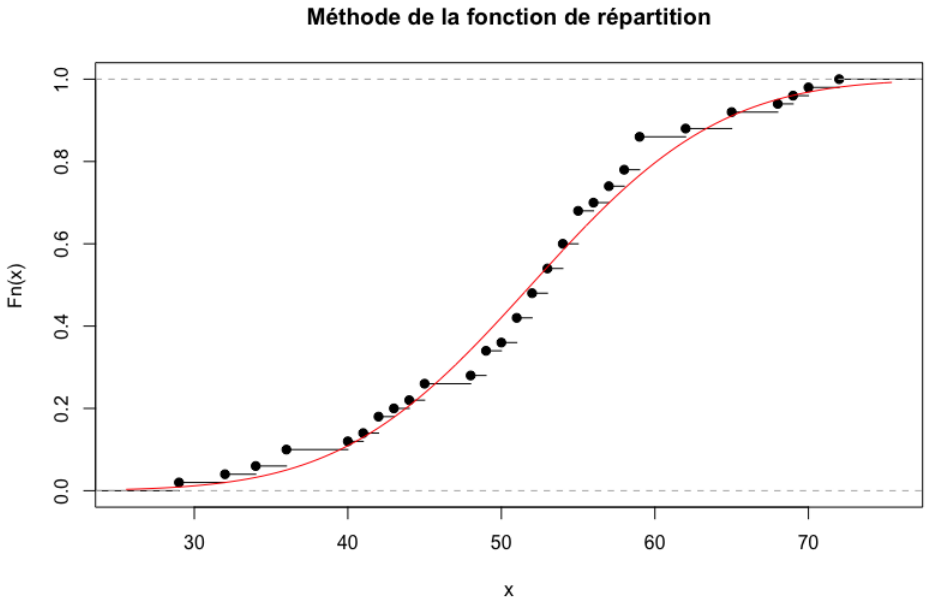
Cela renvoie :



Vu les différences observées, il est difficile de conclure.
— **Méthode de la fonction de répartition :** On propose les commandes :

```
> plot(ecdf(x), main = "Méthode de la fonction de répartition")
> curve(pnorm(x, 51.94, 9.704638), add = TRUE, col="red")
```

Cela renvoie :

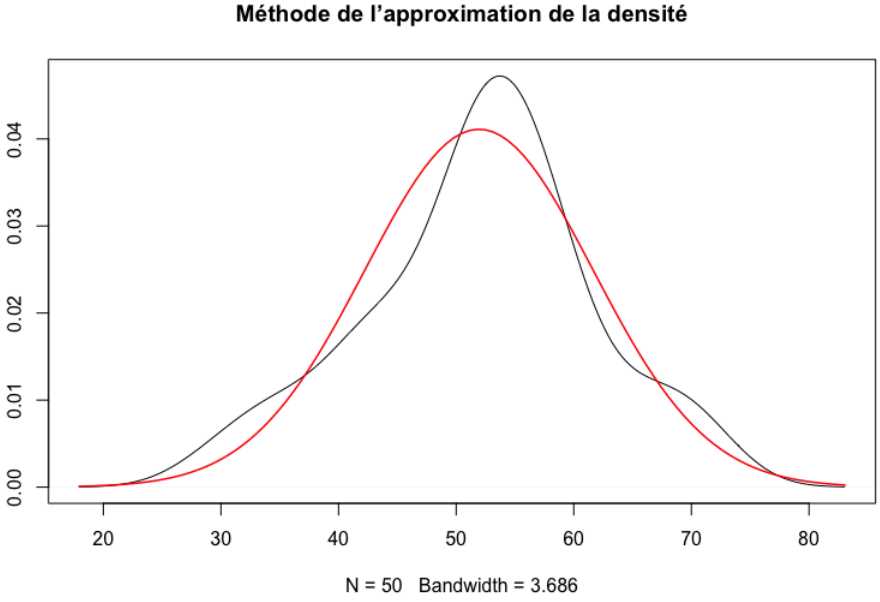


Ici, les différences observées laissent penser que X suit bien une loi normale.

— **Méthode de l'approximation de la densité :** On propose les commandes :

```
> plot(density(x), main = "Méthode de l'approximation de la densité",
+      ylab = "")
> curve(dnorm(x, 51.94, 9.704638), lwd = 1.5, col = "red", add = TRUE)
```

Cela renvoie :

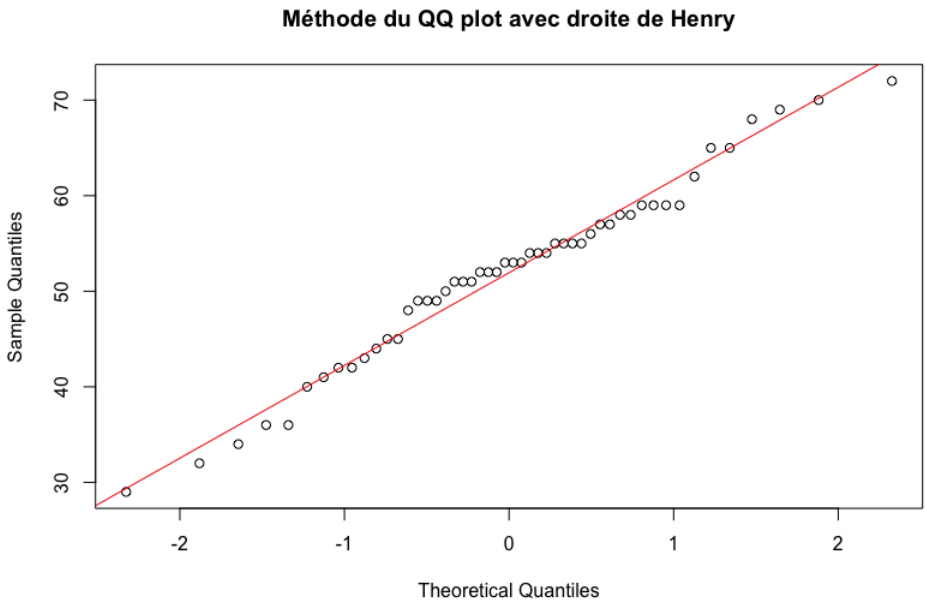


Les informations *"N = 50 Bandwitdth = 3.686"* précisent des quantités utilisées dans l'estimateur de la densité : c'est un estimateur dit "à noyau", lequel utilise une fenêtre (Bandwidth) qui s'ajuste en fonction des données.
De nouveau, les différences observées laissent penser que X suit bien une loi normale.

— **Méthode du QQ plot (cas d'une loi normale) :** On propose les commandes :

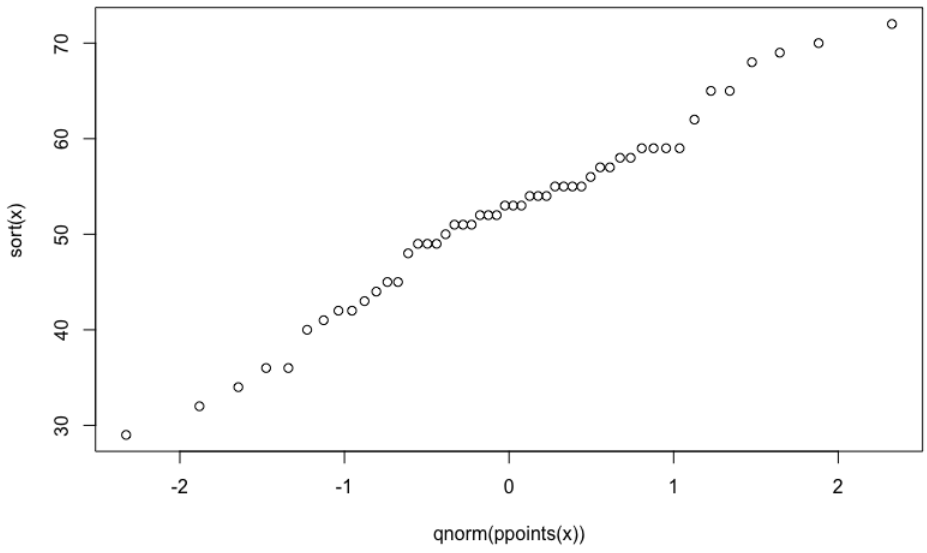
```
> qqnorm(x, main = "Méthode du QQ plot avec droite de Henry")
> a = mean(x) ; b = sd(x);a;b
[1] 51.94
[1] 9.704638
> curve(a + b * x, -6, 6, col = "red", add = TRUE)
```

Cela renvoie :



La droite de henry ajuste bien le nuage de points ; on peut envisager que X suit une loi normale.
À la place de la commande qqnorm, on aurait pu faire le QQ plot à la main :

```
> plot(qnorm(ppoints(x)), sort(x))
```

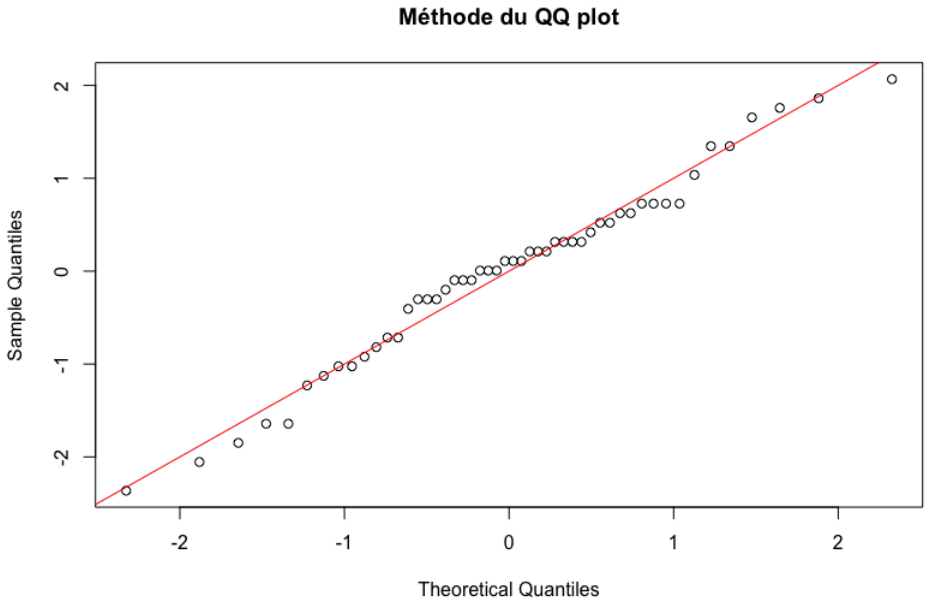


Une autre solution est de "centrer" et "réduire" les données :

$$x_i^* = \frac{x_i - \bar{x}}{s}, i \in \{1, \dots, n\}$$

considérer la loi normale centrée réduite $\mathcal{N}(0;1)$, tracer le QQ plot associé et évaluer son ajustement par la droite $y = x$:

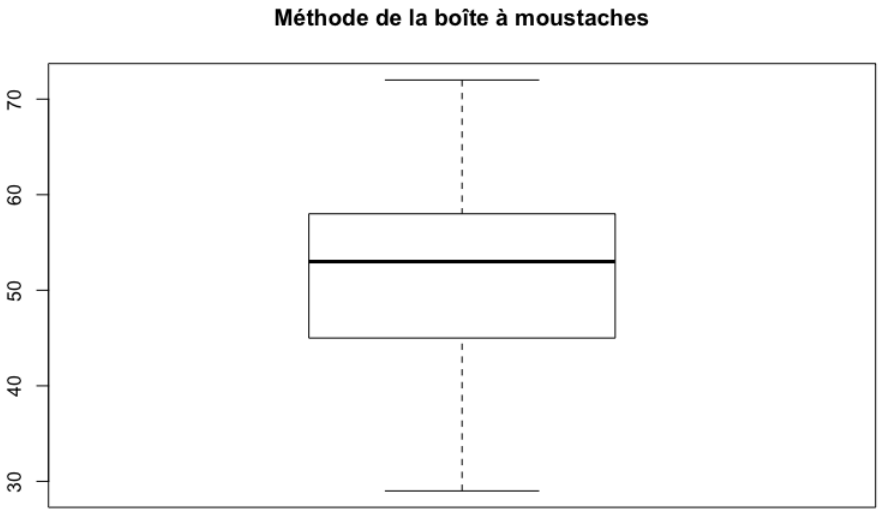
```
> qqnorm(scale(x), main = "Méthode du QQ plot")
> abline(0, 1, col = "red")
```



— **Cas d'une loi normale : méthode de la boîte à moustaches :** On fait :

```
> boxplot(x, main = "Méthode de la boîte à moustaches")
```

Cela renvoie :



On constate qu'aucun point ne se trouve en dehors des "moustaches". D'autre part, la boîte est à peu près symétrique par rapport à la médiane, idem pour les moustaches. On peut envisager que X suit une loi normale.

Conclusion : Toutes les analyses graphiques laissent penser que X suit une loi normale.

2. Dans une station service, un jour donné, les durée du passage en caisse de 49 clients on été mesurés. Les résultats en secondes sont :

0.96	1.45	0.42	3.69	2.58	1.95	1.74	0.01	1.02	1.12
0.17	3.19	0.85	1.27	0.68	3.60	1.23	0.34	0.31	0.16
0.07	0.79	0.02	1.20	0.05	2.09	0.24	5.46	2.57	0.89
0.74	1.67	0.88	2.27	0.22	3.39	0.12	0.06	0.78	0.32
5.79	2.09	0.39	1.82	2.96	0.20	0.08	0.37	2.58	0.30

Soit X la var égale à la durée de passage d'un client. On s'interroge sur le fait que X suit ou non une loi exponentielle.

Solution 1. Par l'énoncé, on observe la valeur de X sur chacun des n individus (clients) d'un échantillon avec $n = 50$: $(x_1; \dots; x_n)$ (avec $x_i \in \mathbb{R}$).
Dire que X suit une loi exponentielle $\varepsilon(\lambda)$, avec $\lambda > 0$ inconnu, signifie qu'elle possède la densité :

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si} \quad x \geq 0 \\ \\ 0 & \text{sinon} \end{cases}$$

Pour préciser la loi exponentielle $\varepsilon(\lambda)$ la plus adaptée à notre contexte, il faut estimer λ à l'aide des données. Comme $\mathbb{E}(X) = 1/\lambda \Leftrightarrow \lambda = 1/\mathbb{E}(X)$, la méthode des moments nous assure qu'une estimation de λ est $1/\bar{x}$. On peut la calculer en faisant :

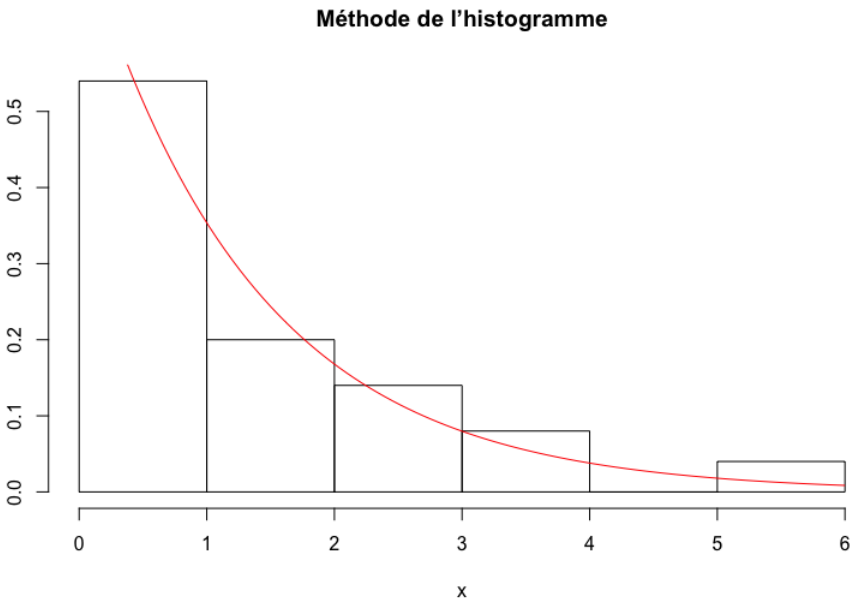
```
> x = c(0.96, 1.45, 0.42, 3.69, 2.58, 1.95, 1.74, 0.01, 1.02, 1.12, 0.17,
+       3.19, 0.85, 1.27, 0.68, 3.60, 1.23, 0.34, 0.31, 0.16, 0.07, 0.79, 0.02,
+       1.20, 0.05, 2.09, 0.24, 5.46, 2.57, 0.89, 0.74, 1.67, 0.88, 2.27, 0.22,
+       3.39, 0.12, 0.06, 0.78, 0.32, 5.79, 2.09, 0.39, 1.82, 2.96, 0.20, 0.08,
+       0.37, 2.58, 0.30)
> 1 / mean(x)
[1] 0.7446016
```

Ainsi, la problématique est la suivante : *est-ce que ces données nous permettent d'affirmer que X ne suit pas la loi $\mathcal{L} = \varepsilon(0 : 7446016)$?*
Nous allons faire une analyse graphique en utilisant les méthodes présentées précédemment.

— **Méthode de l'histogramme :** On propose les commandes :

```
> hist(x, freq = FALSE, main = "Méthode de l'histogramme", ylab = "")
> curve(dexp(x, 0.7446016), add = TRUE, col="red")
```

Cela renvoie :

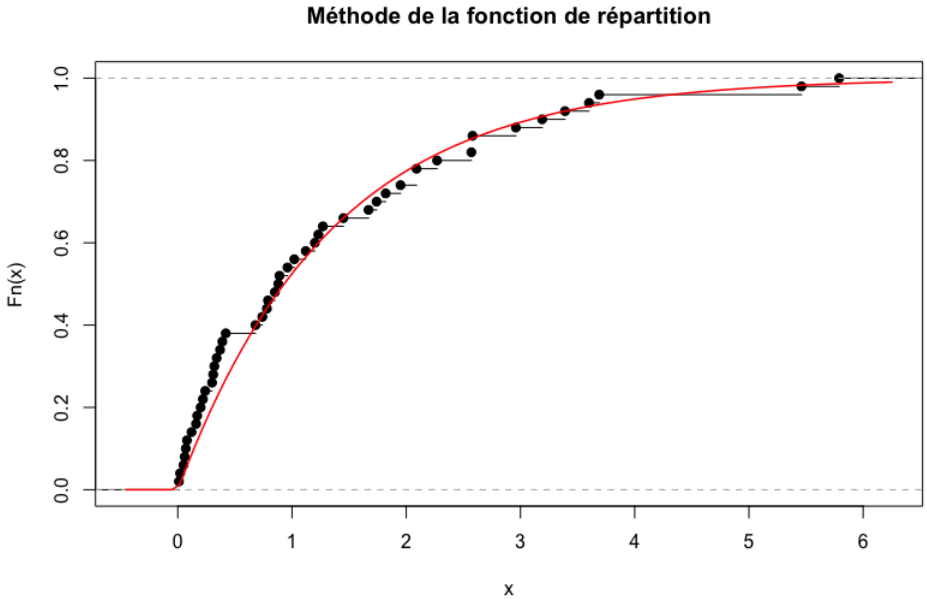


Vu les différences observées, il est vraisemblable que X suit une loi exponentielle.

— **Méthode de la fonction de répartition :** On propose les commandes :

```
> plot(ecdf(x), main = "Méthode de la fonction de répartition")
> curve(pexp(x, 0.7446016), add = TRUE, col="red",lwd=1.5)
```

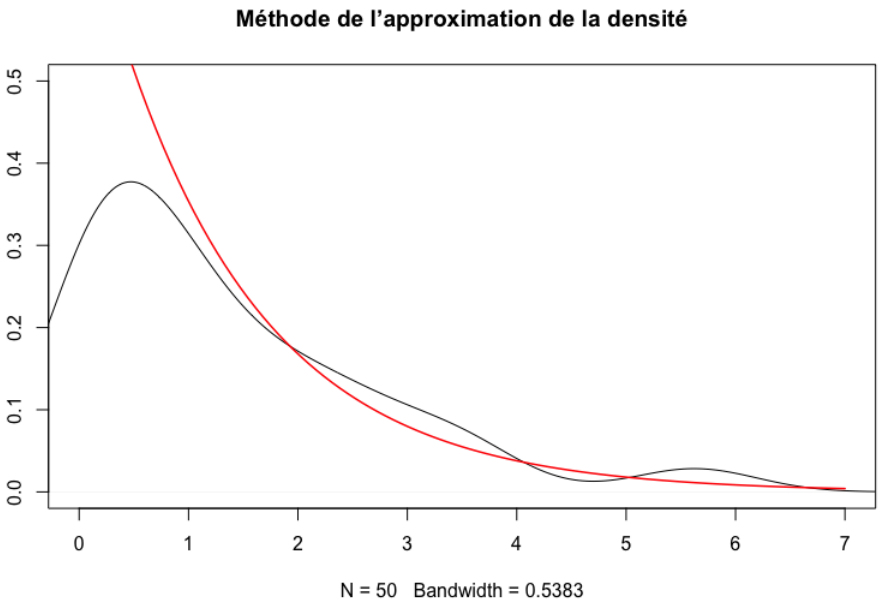
Cela renvoie :



Les différences observées laissent penser que X suit bien une loi exponentielle.

— **Méthode de l'approximation de la densité :** On propose les commandes :

```
> plot(density(x), main = "Méthode de l'approximation de la densité",
+      xlim = c(0, 7), ylim = c(0, 0.5), ylab = "")
> curve(dexp(x, 0.7446016), lwd = 1.5, col = "red", add = TRUE)
```



De nouveau, les différences observées laissent penser que X suit bien la loi exponentielle.

— **Méthode du QQ plot :** On propose les commandes :

```
> plot(qexp(ppoints(x), 0.7446016), sort(x), main = "Méthode du QQ plot",
+      xlab = "")
> abline(0,x, col = "red")
```

La droite diagonale $y = x$ ajuste bien le nuage de points ; on peut envisager que X suit une loi exponentielle.

Conclusion : Toutes les analyses graphiques laissent penser que X suit une loi exponentielle.

1.3 Tests statistiques d'adéquation à une loi

Test du Chi-deux

Données

On observe la valeur d'une *var* X sur chacun des n individus d'un échantillon avec $n \geq 50$. Ces valeurs constituent les données : $x_1; \ldots; x_n$. Elles sont généralement présentées sous la forme d'un tableau classes-effectifs :

Classes	**Effectifs**
C_1	n_1
$\vdots$	$\vdots$
C_k	n_k

Dans ce tableau, pour tout $i \in \{1; : : : ; k\}$, C_i peut être une des valeurs $x_1; \ldots; x_n$, ou un intervalle de valeurs et n_i est le nombre d'individus dont l'observation de X appartient la classe C_i.

Mise en oeuvre

Soit $\mathcal{L}$ une loi de probabilité coïncidant possiblement / étant possiblement en adéquation avec la loi inconnue de X. On considère les hypothèses :

$$H_0 : \text{" } X \text{ suit la loi } \mathcal{L} \text{ " } \textbf{contre } H_1 : \text{" } X \text{ ne suit pas la loi } \mathcal{L} \text{ ".}$$

Pour pouvoir décider du rejet de H_0 :

— Si besoin est, on estime ponctuellement les l paramètres inconnus de la loi $\mathcal{L}$ à l'aide des données, et on considère une *var* R suivant cette loi définie avec les paramètres estimés.

— Si besoin est, on ajuste la première et la dernière classes : C_1 et C_k, de sorte que :

$$\bigcup_{i=1}^{k} C_i = R(\Omega)$$

— Pour tout $i \in \{1; \ldots; k\}$, on calcule la probabilité :

$$p_i = \mathbb{P}(R \in C_i)$$

On vérifie que, pour tout $i \in \{1; \ldots; k\}$, $np_i \geq 5$. Si tel n'est pas le cas, on créé une ou plusieurs nouvelles classes par fusion des anciennes, redéfinissant ainsi le n_i (et le k), jusqu'à obtenir de nouvelles probabilités p_i vérifiant cette hypothèse.

— On calcule :

$$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{n_i^2}{np_i} - n$$

— Soit $K \sim \chi^2(v)$, $v = k - 1 - l$, Alors la p-valeur associée au test du Chi-deux d'adéquation à une loi est :

$$p\text{-valeur} = \mathbb{P}(K \geq \chi^2_{obs})$$

Commandes

En pratique, le test du chi-deux d'adéquation à une loi est surtout utilisé pour une *var* X discrète ; il existe d'autres tests statistiques statistiques "plus puissants" dans le cas où *var* X est à densité.
Lorsqu'aucun paramètre de la loi n'est à estimer, on propose les commandes :

```r
chisq.test(nb, p = proba)$p.value
```

Lorsqu'un ou plusieurs paramètres d'une loi discrète sont à estimer, en introduisant le degré de liberté [2] ajusté *deg*, les commandes donnent la bonne p-valeur :

```r
x2obs = chisq.test(nb, p = proba)$statistic
deg = nbdeclasses - 1 - nbdeparametresestimes
1 - pchisq(x2obs, deg)
```

Si un "Warning message" du type "Chi-squared approximation may be incorrect" apparaît, il faut vérifier que $npi \geq 5$ pour tout $i \in \{1; : : : ; k\}$ à la main et modifier les classes en fonction.
En disposant les données brutes dans un vecteur x , on peut essayer :

```r
library(vcd)
gf = goodfit(x, type = "poisson", method = "MinChisq")
summary(gf)
plot(gf)
```

Exemples

1. La marque Smies produit des bonbons au chocolat de six couleurs. Le responsable de la communication affirme que la proportion de chaque couleur est très précisément de 30% pour le brun (B), 20% pour le jaune (J), 20% pour le rouge (R), 10% pour l'orange (O), 10% pour le vert (V) et 10% pour le doré (D) dans tout échantillon de grande taille. Une expérience réalisée sur un échantillon de 370 bonbons donne les comptages suivants :

Couleur	B	J	R	O	V	D
Nombre de bonbons	84	79	75	49	36	47

Peut-on affirmer, au risque 5%, que le responsable de la communication à tort ?

Solution : Soit X la *var* égale à la couleur d'un bonbon choisi au hasard. Par l'énoncé, on observe la valeur de X sur chacun des n individus (bonbons) d'un échantillon avec $n = 370$: $(x_1; \ldots ; x_n)$. Ces valeurs sont regroupées en $k = 6$ classes : $C_1 = \{B\}$, $C_2 = \{J\}$, $C_3 = \{R\}$,

2. En statistiques le degré de liberté (ddl) désigne le nombre de variables aléatoires qui ne peuvent être déterminées ou fixées par une équation (notamment les équations des tests statistiques).
Par exemple si l'on cherche deux chiffres dont la somme est 12, aucun des deux chiffres ne peut être directement déterminé par la simple équation $X + Y = 12$.
X peut être choisi arbitrairement, mais alors pour Y il n'y a plus le choix. Ainsi, si vous choisissez 11 comme valeur pour X, Y vaut obligatoirement 1. Il y a donc deux variables aléatoires (X, Y), mais un seul degré de liberté.
Une autre définition est fournie par *Walker, H. M.* dans son article *« Degrees of freedom »* publié en 1940 dans Journal of Educational Psychology, (Vol 31(4), Apr 1940, 253-269. doi : 10.1037/h0054588). : *« the number of observations minus the number of necessary relations among these observations »*
Le nombre de degré de liberté est égal au nombre d'observations moins le nombre de relations entre ces observations. On pourrait remplacer l'expression « nombre de relations » par « nombre de paramètres à estimer ».
Un problème est identifié (il présente une solution unique) si le nombre de degrés de liberté est égal à 0.
Si l'on considère une droite ($y = ax + b$), pour définir cette droite nous cherchons à estimer deux paramètres (la pente a et l'origine b). Pour que le problème présente une solution unique il faut remplir la condition *ddl=0*, il nous faut donc 2 observations.
   - La première observation est le Point 1 : $y_1 = ax_1 + b$. Par un point passent une infinité de droites (*ddl=1 observation - 2 paramètres à estimer <0*). Le problème n'a pour le moment pas de solution unique.
   - La seconde observation est le Point 2 : $y_2 = ax_2 + b$, qui vient identifier le problème. Par deux points distincts passe une seule et unique droite (*2 observations - 2 paramètres à estimer = 0 degré de liberté*).

$C_4 = \{O\}$, $C_5 = \{V\}$ et $C_6 = \{D\}$, avec pour effectifs respectifs : $n_1 = 84$, $n_2 = 79$, $n_3 = 75$, $n_4 = 49$, $n_5 = 36$ et $n_6 = 47$. On considère les hypothèses :
— H_0 : "X suit la loi de probabilité décrite par le responsable de la communication"
— H_1 : "X ne suit pas la loi de probabilité décrite par le responsable de la communication".
Soit R une var suivant la loi de probabilité décrite par le responsable de la communication. On a $p_1 = \mathbb{P}(R \in C_1) = 0.3$,$p_2 = \mathbb{P}(R \in C_2) = 0.2$,$p_3 = \mathbb{P}(R \in C_3) = 0.2$,$p_4 = \mathbb{P}(R \in C_4) = 0.1$,$p_5 = \mathbb{P}(R \in C_5) = 0.1$,$p_6 = \mathbb{P}(R \in C_6) = 0.1$.

On considère les commandes :

```
> nb = c(84, 79, 75, 49, 36, 47)
> proba = c(0.3, 0.2, 0.2, 0.1, 0.1, 0.1)
> chisq.test(nb, p = proba)$p.value
[1] 0.01880704
```

Notons qu'aucun "Warning message" n'apparaît ; les conditions d'application du test sont vérifiées. Comme p-valeur $\in]0,01; 0,05]$, le rejet de H_0 est significatif
Par conséquent, au risque 5%, on peut affirmer que X ne suit pas la loi décrite par le responsable de la communication.

2. On souhaite savoir si les entrées à l'hôpital pour une certaine maladie sont réparties au hasard dans l'année ou bien si certains mois sont plus propices à la maladie. On examine le mois d'entrée d'un échantillon de 120 porteurs de la maladie étudiée. Les résultats sont :

Mois d'entrée	1	2	3	4	5	6	7	8	9	10	11	12
Nombre d'entrées	18	16	8	10	6	4	4	9	11	10	12	12

Peut-on affirmer, au risque 1%, que "les entrées ne se font pas au hasard dans l'année" (donc que "certains mois sont plus propices à la maladie") ?

Solution : Soit X la *var* égale au le mois d'entrée à l'hôpital. Par l'énoncé, on observe la valeur de X sur chacun des n individus (porteurs de la maladie) d'un échantillon avec $n = 120$: $(x_1; \dots; x_n)$. Ces valeurs sont regroupées en $k = 12$ classes : $C_1 = \{1\}$, $C_1 = \{2\}$,..., $C_{12} = \{12\}$, avec pour effectifs respectifs : $n_1 = 18$, $n_2 = 16$, ..., $n_{12} = 12$.
Dire que les entrées se font au hasard dans l'année signifie que X suit la loi uniforme $U(\{1; : : : ; 12\})$:

$$\mathbb{P}(X = i) = \frac{1}{12},\ i \in \{1, \dots, 12\}$$

La problématique est la suivante : *est-ce que ces données nous permettent d'affirmer, au risque 1%, que X ne suit pas la loi $\mathcal{L} = U(\{1, \dots, 12\})$?*

On considère alors les hypothèses :
— H_0 : "X suit la loi uniforme $U(\{1, \dots, 12\})$"
— H_1 : "X ne suit pas la loi uniforme $U(\{1, \dots, 12\})$"
Soit R une *var* suivant la loi uniforme $U(\{1, \dots, 12\})$. On a

$$p_i = \mathbb{P}(R \in C_i) = \frac{1}{12},\ \forall i \in \{1, \dots, 12\}$$

On considère les commandes :

```
> nb = c(18, 16, 8, 10, 6, 4, 4, 9, 11, 10, 12, 12)
> proba = rep(1 / 12, 12)
> chisq.test(nb, p = proba)$p.value
[1] 0.04267211
```

Notons qu'aucun "Warning message" n'apparaît ; les conditions d'application du test sont vérifiées. Comme p-valeur $\in]0.01; 0.05]$, le rejet de H0 est significatif.
Toutefois, comme p-valeur > 0.01, au risque 1%, les données ne nous permettent pas d'affirmer que X ne suit pas la loi uniforme $U(\{1, \dots, 12\})$; on ne peut donc rien conclure sur le fait que les entrées ne se font pas au hasard dans l'année.

3. Dans un verger, on étudie le comportement des insectes quand ceux-ci attaquent les fruits. Soit X le caractère qui dénombre le nombre d'attaques d'insectes sur un fruit pris au hasard. Une étude statistique antérieure montre que, si les attaques se font de façon indépendantes les unes des autres, on peut modéliser X comme une *var* suivant une loi de Poisson $\mathcal{P}(\lambda)$ avec λ inconnu. On considère un échantillon de 300 fruits et on compte le nombre d'attaques sur le fruit. Les résultats sont :

Nombre d'attaques	0	1	2	3	4	5	6	7
Nombre de fruits attaqués	60	105	65	47	15	4	3	1

Peut-on dire que le comportement des insectes est significativement grégaire ? (on dit que le comportement des insectes est grégaire quand chacun d'entre eux a tendance à se comporter comme le voisin et à attaquer le même fruit ; dans ce cas, X ne suit pas une loi de Poisson).

Solution Par l'énoncé, on observe la valeur de X sur chacun des n individus (fruits) d'un échantillon avec $n = 300$: $(x_1; \dots; x_n)$ (avec $x_i \in \mathbb{N}$). Ces valeurs sont regroupées en $k = 8$ classes : $C_1 = \{0\}$, $C_2 = \{1\}$,..., $C_8 = \{7\}$ avec pour effectifs respectifs : $n_1 = 60$, $n_2 = 105$,..., $n_8 = 1$.
La problématique est la suivante : *est-ce que ces données nous permettent d'affirmer, a moins au risque 5%, que X ne suit pas une loi de Poisson ?*
On considère les hypothèses :
— H_0 : "X suit une loi de Poisson"
— H_1 : "X ne suit pas une loi de Poisson".
Soit R une var suivant la loi de Poisson $\mathcal{P}(\lambda)$ avec $\lambda > 0$ inconnu. On a $R(\Omega) = \mathbb{N}$.
Comme $\bigcup_{i=1}^{8} C_i = \{0; \dots; 7\} \neq \mathbb{N}$, on ajuste la dernière classe comme : $C_8 = \{7; 8; \dots; \infty\}$ (de sorte à ce que $\bigcup_{i=1}^{8} = R(\Omega)$. Le paramètre λ étant inconnu, il faut l'estimer à l'aide des données. Comme $\mathbb{E}(X) = \lambda$, la méthode des moments nous assure qu'une estimation de λ est la moyenne x que l'on peut calculer en faisant :

```
> nb = c(60, 105, 65, 47, 15, 4, 3, 1)
> lambda = sum((0:7) * nb) / 300
> lambda
[1] 1.603333
```

Cette estimation sera prise en compte dans la suite. Pour tout $i \in \{1; \dots; 7\}$, on a

$$p_i = \mathbb{P}(R \in C_i) = \mathbb{P}(R = i - 1) = e^{-1.603333} \frac{1.603333^{i-1}}{(i-1)!}$$

et

$$p_8 = \mathbb{P}(R \in C_8) = \mathbb{P}(R \geq 7) = \sum_{i=1}^{\infty} e^{-1.603333} \frac{1.603333^{i}}{i!}$$

On peut obtenir ces probabilités avec les commandes :

```
> proba = c(dpois(0:6, lambda), 1 - ppois(6, lambda))
> proba
[1] 0.201224650 0.322630189 0.258641868 0.138229709 0.055407075 0.017767202
[7] 0.004747791 0.001351515
```

Dans un premier temps, on propose de mettre en oeuvre le test du Chi-deux en faisant :

```
> chisq.test(nb, p = proba)$p.value
[1] 0.4732928
Warning message:
In chisq.test(nb, p = proba) :
  l'approximation du Chi-2 est peut-être incorrecte
```

On dénombre alors deux problèmes :

— le logiciel n'a pas pris en compte le fait que λ a été estimé (il ne peut pas le savoir)

— il y a un "Warning message" nous avertissant que l'hypothèse : $np_i \geq 5$ pour tout $i \in \{1; \dots ; 8\}$ n'est peut-être pas vérifiée.

Étudions ce dernier point :

```
> proba
[1] 0.201224650 0.322630189 0.258641868 0.138229709 0.055407075 0.017767202
[7] 0.004747791 0.001351515
> 300 * proba
[1] 60.3673950 96.7890567 77.5925604 41.4689128 16.6221226  5.3301606
[7]  1.4243374  0.4054545
```

Comme $np_7 < 5$ et $np_8 < 5$, nous allons fusionner les classes C_6, C_7 et C_8, formant ainsi une nouvelle dernière classe : $C_6 = \{5; 6; \dots\}$ avec pour effectif : $n_6 = 4 + 3 + 1 = 8$. Il y a désormais $k = 6$ classes. On vérifie alors que l'hypothèse est vérifiée avec cette nouvelle configuration :

```
> nb2 = c(60, 105, 65, 47, 15, 8)
> proba2 = c(dpois(0:4, lambda), 1 - ppois(4, lambda))
> proba2
[1] 0.20122465 0.32263019 0.25864187 0.13822971 0.05540708 0.02386651
> 300 * proba2
[1] 60.367395 96.789057 77.592560 41.468913 16.622123  7.159953
```

Aucune valeur ne dépasse 5. Pour avoir la p-valeur associée au test du Chi-deux en prenant en compte le fait que l'on a estimé un paramètre, on considère le degré de liberté :

$$\mu = k - 1 - l = 6 - 1 - 1 = 4$$

On fait :

```
> x2obs = chisq.test(nb2, p = proba2)$statistic
> deg = 4
> 1 - pchisq(x2obs, deg)
X-squared
0.4427608
```

On obtient alors la vraie p-valeur associée au test du Chi-deux (il ne faut pas faire attention au "X-squared" et notons aussi qu'aucun "Warning message" n'apparaît). Comme p-valeur > 0.05, les données ne nous permettent pas de rejeter H_0. Ainsi, on ne peut pas rejeter l'hypothèse selon laquelle les insectes n'ont pas un comportement grégaire.

Test de Kolmogorov-Smirnov

Contexte

On observe la valeur d'une *var* X sur chacun des n individus d'un échantillon. Ces valeurs constituent les données : $x_1; \dots ; x_n$. On considère les hypothèses :

$$H_0 \text{ : " } X \text{ suit la loi } \mathcal{L} \text{ " } \textbf{contre } H_1 \text{ : " } X \text{ ne suit pas la loi } \mathcal{L} \text{ ".}$$

Soient $F_n(x)$ la fonction de répartition empirique associée aux données et $F(x)$ la fonction de répartition associé à la loi $\mathcal{L}$. L'idée du test de Kolmogorov-Smirnov est que plus $F_n(x)$ diffère de $F(x)$, plus le rejet de H_0 est significatif (idée similaire à l'analyse graphique de la méthode de la fonction de répartition). La p-valeur du test de Kolmogorov-Smirnov utilise la statistique de test observée :

$$d_{obs} = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

et une loi de probabilité non-usuelle que l'on arrive à évaluer.

En pratique, le test de Kolmogorov-Smirnov est surtout utilisé pour une *var* X à densité. Il est plus **puissant que le test du Chi-deux** quand n est petit.
Les commandes associées sont :

```
ks.test(x, "pnorm")$p.value
```

Lorsqu'un ou plusieurs paramètres de la loi sont à préciser, on fait, par exemple :

```
ks.test(x, "pexp", lambda)$p.value
```

Exemple

On mesure les durées de vie de 20 ampoules d'un même type. Les résultats, en heures, sont :

673	389	1832	570	522	2694	3683	644	1531	2916
1069	3145	2268	3574	791	1418	649	3344	1153	3922

Soit X la *var* égale à la durée de vie en heures d'une ampoule de ce type.
Est-ce que l'on peut affirmer, au risque 5%, que X ne suit pas la loi exponentielle $\mathcal{E}(1/1850)$?

Solution Par l'énoncé, on observe la valeur de X sur chacun des n individus (ampoules) d'un échantillon avec $n = 20$: $(x_1; \dots; x_n)$ (avec $x_i \in \mathbb{R}$).
On considère les hypothèses :

— H_0 : "X suit la loi exponentielle $\mathcal{E}(1 = 1850)$"
— H_1 : "X ne suit pas la loi exponentielle $\mathcal{E}(1 = 1850)$"

On peut utiliser le test de Kolmogorov-Smirnov. On fait :

```
> x = c(673, 389, 1832, 570, 522, 2694, 3683, 644, 1531, 2916, 1069, 3145,
+       2268, 3574, 791, 1418, 649, 3344, 1153, 3922)
> ks.test(x, "pexp", 1 / 1850)$p.value
[1] 0.3774748
```

Comme p-valeur > 0.05, les données ne nous permettent pas de rejeter H_0. Ainsi, on ne peut pas rejeter l'hypothèse selon laquelle X suit la loi exponentielle $\mathcal{E}(1/1850)$.

Test de Shapiro-Wilk

Contexte

On observe la valeur d'une *var* X sur chacun des n individus d'un échantillon. Ces valeurs constituent les données : $x_1; \dots; x_n$. On cherche à montrer que X ne suit pas une loi normale, ce qui mettrait en défaut une hypothèse cruciale pour de nombreux outils statistiques comme des intervalles de confiance (T-IntConf,...) et des tests statistiques (T-Test, test du coefficient de corrélation, ANOVA ...).
On considère donc les hypothèses :

— H_0 : "X suit une loi normale $\mathcal{N}$"
— H_1 : "X ne suit pas une loi normale $\mathcal{N}$".

On peut alors utiliser le test du Chi-deux ou le test de Kolmogorov-Smirnov avec $\mathcal{L}$ = loi normale $\mathcal{N}$. Toutefois, dans ce cas particulier, il est fortement conseillé d'utiliser le test de Shapiro-Wilk plus puissant. Celui-ci utilise la statistique de test observée :

$$w_{obs} = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{(n-1)s^2}$$

où les valeurs $a_1; \ldots; a_n$ sont calculés à partir du vecteur des moyennes et de la matrice de covariance des statistiques d'ordre de *n var iid* suivant une loi normale et une loi de probabilité non-usuelle que l'on arrive à évaluer.
Les commandes associées sont :

```
shapiro.test(x)$p.value
```

Remarque : Le test de Shapiro-Wilk appartient à la grande famille des "tests de normalité". Il en existe un grands nombres, parmi lesquels : le test de Lilliefors, le test de Anderson-Darling, le test de D'Agostino et le test de Jarque-Bera. Toutefois, s'il fallait en retenir qu'un, ca serait le test de Shapiro-Wilk.

Exemple

On fait passer à 50 adolescents le test psychologique de Rorschach. Les temps de passation en minutes du test sont :

43	48	65	55	51	51	44	51	59	62
45	53	55	55	49	34	52	69	45	54
59	36	36	29	52	59	41	58	54	55
72	53	52	49	57	42	70	58	42	53
57	68	40	65	54	49	32	56	50	59

Soit X la var qui à un adolescent associe son temps de passation au test. *Peut-on affirmer, au risque* 5%*, que X ne suit pas une loi normale ?*

Solution Par l'énoncé, on observe la valeur de X sur chacun des n individus (adolescents) d'un échantillon avec $n = 50$: $(x_1; \ldots; x_n)$ (avec $x_i \in \mathbb{R}$). On considère les hypothèses :
— H_0 : "X suit une loi normale"
— H_1 : "X ne suit pas une loi normale".

On peut utiliser le test de Shapiro-Wilk. On fait :

```
> x = c(43, 48, 65, 55, 51, 51, 44, 51, 59, 62, 45, 53, 55, 55, 49, 34, 52,
+       69, 45, 54, 59, 36, 36, 29, 52, 59, 41, 58, 54, 55, 72, 53, 52, 49, 57, 42,
+       70, 58, 42, 53, 57, 68, 40, 65, 54, 49, 32, 56, 50, 59)
> shapiro.test(x)$p.value
[1] 0.4801461
```

Comme p-valeur $> 0,05$, on ne rejette pas H_0 ; l'hypothèse selon laquelle X suit une loi normale n'est pas rejetée.

1.4 Exercices

1.4.1 Enoncés

Exercice 1

On lance 100 fois un dé à 6 faces numérotées de 1 à 6. On obtient les résultats suivants :

Numéro	1	2	3	4	5	6
Nombre de fois	18	23	19	12	11	15

Peut-on affirmer que le dé est truqué ? (on fera une analyse graphique convenable, puis un test statistique adapté au risque 5%).

Exercice 2

On indique dans le tableau suivant le nombre de fois qu'un chiffre apparaît dans les 608 premières décimales de π :

Chiffre	0	1	2	3	4	5	6	7	8	9
Nombre de fois	60	62	67	68	64	56	62	44	58	67

Peut-on affirmer, au risque 5%, que les décimales ne sont pas équiréparties ?

Exercice 3

Lindström, spécialiste de la génétique et de l'hybridation du maïs, a croisé deux types récessifs de maïs : le type vert-zébré et le type doré. Si les lois de la génétique sont respectées obtient :

— "vert" avec la probabilité $9/16$
— "doré" avec la probabilité $3/16$
— "vert-zébré" avec la probabilité $3/16$
— "doré-vert-zébré" avec la probabilité $1/16$

On effectue 1301 croisements. On obtient les résultats suivants :

Type	"vert"	"doré"	"vert-zébré"	"doré-vert-zébré"
Nombre defois	773	231	238	59

Peut-on dire que les lois de la génétique ne sont pas respectées ? (on fera une analyse graphique convenable, puis un test statistique adapté au risque 5%).

Exercice 4

L'agence immobilière Centurion a étudié le nombre de biens vendus par agent le mois de Juin 2016. Les résultats obtenus sont :

Nombre de biens vendus	0	1	2	≥ 3
Nombre d'agents	15	18	11	8

Soit X la *var* égale au nombre de biens vendu par un agent en Juin 2016.*Peut-on affirmer que X ne suit pas la loi de Poisson $\mathcal{P}(2)$?* (on fera une analyse graphique convenable, puis un test statistique adapté au risque 5%).

Exercice 5

Dans le garage AutoLover, sur une période de 320 jours, on a relevé le nombre journalier d'accidents du travail. Les résultats obtenus sont :

Nombre d'accidents journalier	0	1	2	3	4	5	6	7
Nombre de jours	65	110	70	48	16	5	4	2

On peut modéliser le nombre journalier d'accidents du travail dans ce garage par une var X. *Peut on affirmer, au risque* 5%*, que* X *ne suit pas une loi de Poisson ?* (on fera une analyse graphique convenable, puis un test statistique adapté au risque 5%).

Exercice 6

On mesure les durées de vie en heures de 7 appareils. Les résultats obtenus sont :

Appareil	1	2	3	4	5	6	7
Durée de vie	145	110	170	48	116	95	74

On peut modéliser la durée de vie en heures d'un appareil par une var X. Peut-on affirmer que X, au risque 5%, que ne suit pas une loi exponentielle $\varepsilon(0.01)$?

Exercice 7

Dans un magasin, un jour donnée, on mesure les temps d'attente en minutes entre 2 clients à une caisse. Les résultats obtenus sont :

25.12	12.36	24.35	12.19	5.27	18.35	19.11	27.08	21.09	17.19	8.45	13.27	15.17

On peut modéliser le temps d'attente entre 2 clients par une var X. Peut-on affirmer que X, au risque 5%, que ne suit pas une loi exponentielle $\varepsilon(1/\bar{x})$, $\bar{x}$ où désigne la moyenne des valeurs obtenues ?

Exercice 8

On étudie l'accroissement de poids en kilogrammes chez des pourceaux pendant une période de 15 jours. On prélève au hasard 50 pourceaux. Les résultats, sous la forme d'un tableau classes-effectifs, sont :

Classes	**Effectifs**
]0,4]	2
]4,8]	5
]8,12]	12
]12,16]	14
]16,20]	11
]20,24]	5
]24,28]	1

Soit X la var égale à l'accroissement de poids en kilogrammes d'un pourceau pendant 15 jours. À l'aide d'une analyse graphique, montrer qu'il est vraisemblable que X suit une loi normale.

Exercice 9

On pèse 20 plaquettes de beurre pris au hasard dans une production normande. Les résultats, en grammes, sont :

247.0	247.8	250.2	251.3	251.9	249.4	248.8	247.1	255.0	247.0
254.8	244.8	250.7	250.7	252.6	251.1	254.1	249.2	252.0	254.0

On suppose que le poids en grammes d'une plaquette de beurre de cette production peut être modélisé par une var X. *Peut-on affirmer que X suit une loi normale ?* (on fera une analyse graphique convenable, puis un test statistique adapté au risque 5%).

Exercice 10

Une ferme de Bay of Plenty en Nouvelle-Zélande produit des kiwis. On pèse 16 kiwis choisis au hasard dans cette ferme. Les résultats, en grammes, sont :

247.0	247.8	250.2	251.3	251.9	249.4	248.8	247.1	255.0	247.0
254.8	244.8	250.7	250.7	252.6	251.1	254.1	249.2	252.0	254.0

Soit X la var égale au poids en grammes d'un kiwi. *Peut-on affirmer que X suit une loi normale ?* (on fera une analyse graphique convenable, puis un test statistique adapté au risque 5%).

Exercice 11

La pression artérielle systolique est la pression maximale du sang dans les artères au moment de la contraction du coeur. Celle-ci a été mesurée pour 29 individus de différents âges. Ainsi, pour chacun d'entre eux, on dispose :

— de leur pression systolique en mmHg,
— de leur âge en années

On modélise ces deux variables comme des var Y et X_1. Le jeu de données "pression" est disponible ici : lien.

1. Mettre le jeu de données sous la forme d'une data frame w, puis attacher les noms des colonnes.
2. Peut-on affirmer, au risque 5%, que Y ne suit pas une loi normale ?
3. Peut-on affirmer, au risque 5%, que X1 ne suit pas une loi normale ?
4. Reproduire et comprendre l'enjeu des commandes suivantes :

```
par(mfrow = c(1, 2))
qqnorm(scale(X1))
abline(0, 1, col = "red")
qqnorm(scale(Y))
abline(0, 1, col = "blue")
```

Exercice 12

Soient X et Y deux *var* indépendantes telles que X suit la loi de Poisson $\mathcal{P}(5)$ et Y suit la loi de Poisson $\mathcal{P}(3)$. Alors on sait que $Z = X + Y$ suit la loi de Poisson $\mathcal{P}(5+3)$.
Illustrer ce résultat en simulant des var et en utilisant la commande qqplot.

Exercice 13

Soient X_1 et X_2 deux var indépendantes. Illustrer les résultats ci-dessous avec la commande qqplot :

$X_i \sim$	$\varepsilon(\lambda)$	$\Gamma(m_i, \lambda)$	$\mathcal{N}(\mu_i, \sigma_i^2)$	$\chi^2(\nu_i)$
$X_1 + X_2 \sim$	$\Gamma(2, \lambda)$	$\Gamma(m_1 + m_2, \lambda)$	$\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$	$\chi^2(\nu_1 + \nu_2)$
Numérique	$\lambda = 3.8$	$m_1 = m_2 = 4.2, \lambda = 2.1$	$\mu_1 = \mu_2 = 1.6, \sigma_1 = \sigma_2 = 1.5$	$\nu_1 = \nu_2 = 3.2$

Exercice 14

Soient X et Y deux var indépendantes. Illustrer les résultats ci-dessous avec la commande qqplot :

— Caractérisation de la loi du chi-deux $\chi^2(2)$:
Si $X \sim \mathcal{N}(0,1)$ et $Y \sim \mathcal{N}(0,1)$ alors

$$X^2 + Y^2 \sim \chi^2(2)$$

— Caractérisation de la loi de Student $\mathcal{T}(\nu)$:
Si $X \sim \mathcal{N}(0,1)$ et $Y \sim \chi^2(\mu)$ alors

$$\frac{X}{\sqrt{\frac{Y}{\nu}}} \sim \mathcal{T}(\nu)$$

Prendre $\nu = 3.9$.

— Caractérisation de la loi de Fisher $\mathcal{F}(\nu_1, \nu_2)$

$$\frac{\frac{X}{\nu_1}}{\frac{Y}{\nu_2}} = \frac{\nu_2 X}{\nu_1 Y} \sim \mathcal{F}(\nu_1, \nu_2)$$

Prendre $(\nu_1, \nu_2) = (2,1;8,3)$.

Exercice 15

On considère les commandes :

```
> x = rnorm(100)
> a = numeric()
> for (i in 1:6) {
+   a[i] = ks.test(x, "pnorm", 0, 1 + (i - 1) / 10)$p.value
+ }
> a
[1] 0.86547276 0.52803787 0.25056642 0.10861385 0.04521945 0.01858058
```

Commenter ces résultats numériques.

La grande philosophie des tests statistiques, c'est de les utiliser pour espérer écarter les hypothèses qui contrediraient la normalité (ou autre) d'une loi

1.4.2 Solutions

Exercice 1

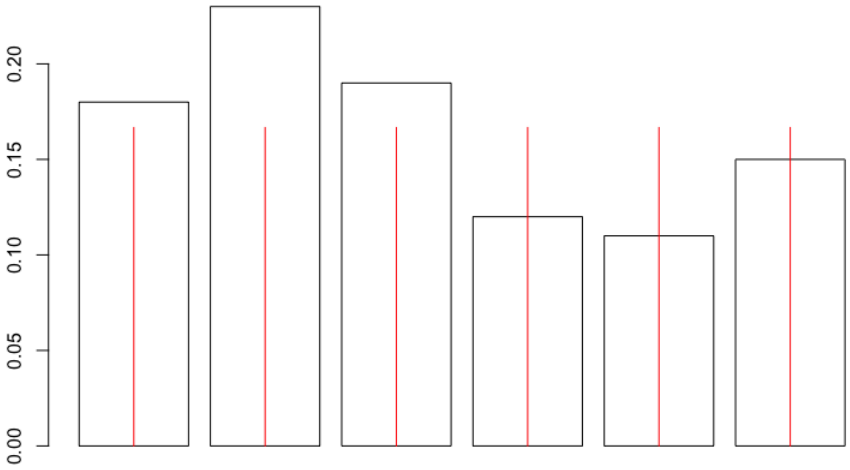
Soit X la *var* égale au numéro affiché par le dé après un lancer. Par l'énoncé, on observe la valeur de X sur chacun des n individus (dés) d'un échantillon avec $n = 100$: $(x_1; \dots; x_n)$ (avec $x_i \in \{1; \dots; 6\}$). On forme alors un vecteur des effectifs $(n_1; n_2; \dots; n_6) = (18; 23; \dots; 15)$.
Dire que le dé n'est pas truqué signifie que X suit la loi uniforme $U(\{1; \dots; 6\})$:

$$\mathbb{P}(X = i) = \frac{1}{6},\ i \in \{1, \dots, 6\}$$

La problématique est la suivante : *est-ce que les données nous permettent d'affirmer que X ne suit pas la loi* $U(\{1; \dots; 6\})$?

Ainsi, pour une analyse graphique, on propose les commandes :

```
> nb = c(18, 23, 19, 12, 11, 15)
> bar = barplot(nb / 100, col = "white")
> points(bar, rep(1 / 6, 6), type = "h",col="red")
```



Il est difficile de conclure au vu des différences observées.
On peut utiliser le test du Chi-deux d'adéquation à une loi pour y voir plus clair. On considère alors les hypothèses :

— H_0 : "X suit la loi uniforme $U(\{1; \dots; 6\})$"
— H_1 : "X ne suit pas la loi uniforme $U(\{1; \dots; 6\})$".

Les valeurs sont regroupées en $k = 6$ classes : $C_1 = \{1\}$, $C_2 = \{2\}$,..., $C_6 = \{6\}$.
On considère les commandes :

```
> nb
[1] 18 23 19 12 11 15
> proba = rep(1 / 6, 6)
```

```
> chisq.test(nb, p = proba)$p.value
[1] 0.2757299
```

Notons qu'aucun "Warning message" n'apparaît ; les conditions d'application du test sont vérifiées.
Comme p-valeur $> 0,05$, les données ne nous permettent pas de rejeter H_0. Ainsi, on ne peut pas affirmer que le dé est truqué.

Exercice 2

Soit X la var égale au chiffre affiché par une décimale (que l'on suppose donc aléatoire).
Par l'énoncé, on observe la valeur de X sur chacun des n individus (décimales) d'un échantillon avec $n = 608$: $(x_1; \dots ; x_n)$ (avec $x_i \in \{0; \dots ; 9\}$). On forme alors un vecteur des effectifs $(n_1; n_2; \dots ; n_{10}) = (60; 62; : : : ; 67)$. Dire que les décimales sont équiréparties signifie que X suit la loi uniforme $U(\{0; \dots ; 9\})$:

$$\mathbb{P}(X = i) = \frac{1}{10},\ i \in \{1, \dots , 9\}$$

La problématique est la suivante : *est-ce que les données nous permettent d'affirmer que X ne suit pas la loi* $U(\{0; \dots ; 9\})$?

— H_0 : "X suit la loi uniforme $U(\{1; \dots ; 9\})$"
— H_1 : "X ne suit pas la loi uniforme $U(\{1; \dots ; 9\})$".
On peut utiliser le test du Chi-deux d'adéquation à une loi. Les valeurs sont regroupées en k $= 10$ classes : $C_1 = \{0\}$, $C_2 = \{1\}$,..., $C_{10} = \{9\}$.
On considère les commandes :

```
> nb = c(60, 62, 67, 68, 64, 56, 62, 44, 58, 67)
> proba = rep(1 / 10, 10)
> chisq.test(nb, p = proba)$p.value
[1] 0.585888
```

Notons qu'aucun "Warning message" n'apparaît ; les conditions d'application du test sont vérifiées.
Comme p-valeur $> 0,05$, les données ne nous permettent pas de rejeter H_0. Ainsi, au risque 5%, on ne peut pas affirmer que le décimales de π soient équiréparties.

Exercice 3

Soit X la var égale au type de maïs obtenu avec un croisement. On utilise le codage :

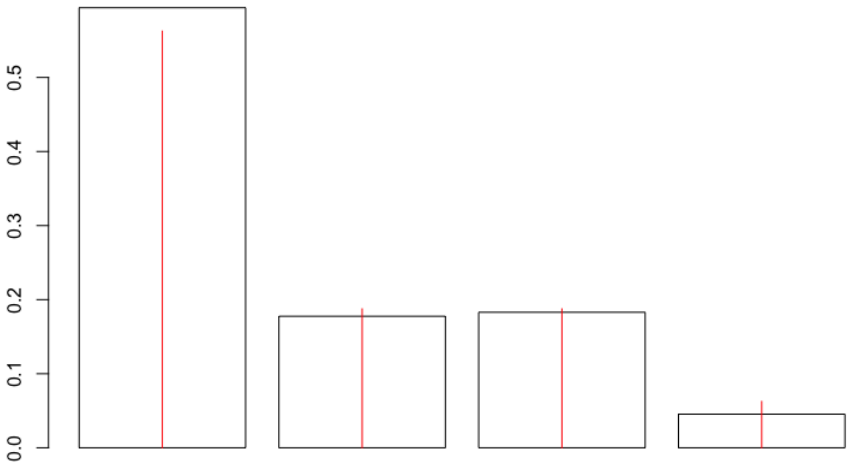
Type	"vert"	"doré"	"vert-zébré"	"doré-vert-zébré"
Nombre defois	0	1	2	3

Par l'énoncé, on observe la valeur de X sur chacun des n individus (croisements) d'un échantillon avec $n = 1301$: $(x_1; \dots ; x_n)$ (avec $x_i \in \{0; 1; 2; 3\}$). On forme alors un vecteur des effectifs $(n_1; n_2; n_3; n_4) = (773; 231; 238; 59)$. Dire que les lois de la génétique sont respectées signifie que X suit la loi décrite dans l'énoncé :

Type	"vert"	"doré"	"vert-zébré"	"doré-vert-zébré"
Nombre de fois	9/16	3/16	3/16	1/16

La problématique est la suivante : *est-ce que les données nous permettent d'affirmer que X ne suit pas la loi décrite dans l'énoncé ?*
Ainsi, pour une analyse graphique, on propose les commandes :

```
> nb = c(773, 231, 238, 59)
> bar = barplot(nb / 1301, col = "white")
> points(bar, c(9 / 16, 3 / 16, 3 / 16, 1 / 16), type = "h",col="red")
```



Il est difficile de conclure au vu des différences observées.
On peut utiliser le test du Chi-deux d'adéquation à une loi pour y voir plus clair.
On considère alors les hypothèses :

— H_0 : "X suit la loi décrite dans l'énoncé"
— H_1 : "X ne suit pas la loi décrite dans l'énoncé".

Les valeurs sont regroupées en $k = 4$ classes : $C_1 = \{1\}$, $C_2 = \{2\}$, $C_3 = \{3\}$ et $C_4 = \{4\}$. On considère les commandes :

```
> nb
[1] 773 231 238  59
> proba = c(9 / 16, 3 / 16, 3 / 16, 1 / 16)
> chisq.test(nb, p = proba)$p.value
[1] 0.02589168
```

Notons qu'aucun "Warning message" n'apparaît ; les conditions d'application du test sont vérifiées.
Comme p-valeur $\in]0,01; 0,05]$, le rejet de H_0 est significatif. Par conséquent, on peut affirmer, au risque 5%, que X ne suit pas la loi décrite dans l'énoncé ; les lois de la génétique ne sont pas respectées.

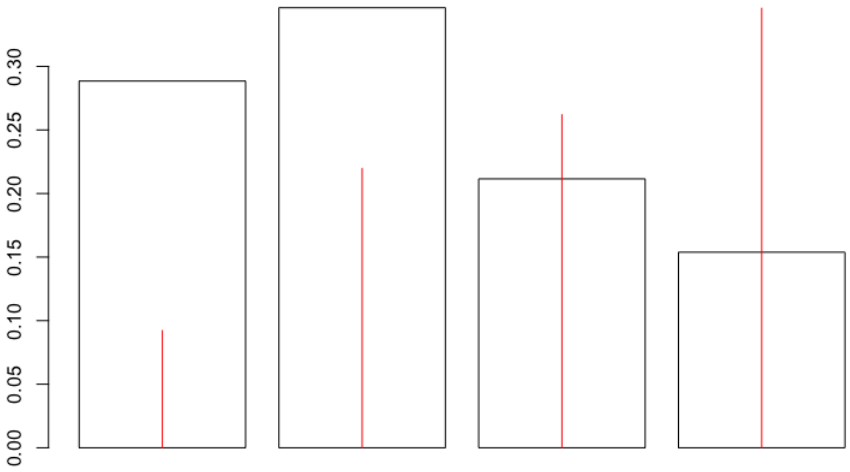
Exercice 4

Par l'énoncé, on observe la valeur de X sur chacun des n individus (agents) d'un échantillon avec $n = 52$: $(x_1; \dots ; x_n)$ (avec $x_i \in \mathbb{N}$). On forme alors un vecteur des effectifs $(n_1; n_2; n_3; n_4) = (15; 18; 11; 8)$.
Dire que X suit la loi de Poisson $\mathcal{P}(2)$ signifie que

$$\mathbb{P}(X = i) = e^{-2}\frac{2^i}{i!},\ i \in \mathbb{N}$$

La problématique est la suivante : *est-ce que ces données nous permettent d'affirmer que X ne suit pas la loi $\mathcal{P}(2)$?* Pour une analyse graphique, on propose les commandes :

```
> nb = c(15, 18, 11, 8)
> bar = barplot(nb / 52, col = "white")
> lambda = sum((0:4) * nb) / 52;lambda
[1] 2.384615
> prob = c(dpois(0:2, lambda), 1 - ppois(2, lambda))
> points(bar, prob, type = "h",col="red")
```



Les différences observées laissent penser que X ne suit pas la loi de Poisson $\mathcal{P}(2)$. Confirmons cela avec le test du Chi-deux d'adéquation à une loi. On considère alors les hypothèses :

— H_0 : "X suit la loi de Poisson $\mathcal{P}(2)$"

— H_1 : "X ne suit pas la loi de Poisson $\mathcal{P}(2)$".

Les données sont sont regroupées en $k = 4$ classes : $C_1 = \{0\}$, $C_2 = \{1\}$,$C_3 = \{2\}$, $C_4 = \{3, 4, \dots, \infty\}$.
Soit R une var suivant la loi de Poisson $\mathcal{P}(2)$. On a $R(\Omega)) = \mathbb{N}$ et $\bigcup_{i=1}^{4} C_i = \mathbb{N} = R(\Omega)$ (il n'y a pas d'ajustement à faire). On met en oeuvre le test du Chi-deux en faisant :

```
> nb
[1] 15 18 11  8
> prob
[1] 0.09212441 0.21968127 0.26192767 0.42626665
> chisq.test(nb, p = prob)$p.value
[1] 1.159166e-07
```

Comme p-valeur $\in]0.001; 0.01]$, le rejet de H_0 est très significatif. Par conséquent, au risque 5%, on peut affirmer que X ne suit pas la loi de Poisson $\mathcal{P}(2)$.

Exercice 5

Par l'énoncé, on observe la valeur de X sur chacun des n individus (jours) d'un échantillon avec $n = 320$: $(x_1; \dots; x_n)$ (avec $x_i \in \mathbb{N}$). Ces valeurs sont regroupées en $k = 8$ classes : $C_1 = \{0\}$, $C_2 = \{1\}$,..., $C_8 = \{7\}$ avec pour effectifs respectifs : $n_1 = 65$, $n_2 = 110$,..., $n_8 = 2$.
On considère les hypothèses :

— H_0 : "X suit une loi de Poisson"

— H_1 : "X ne suit pas une loi de Poisson"

On peut alors utiliser le test du Chi-deux d'adéquation à une loi. Soit R une var suivant la loi de Poisson $\mathcal{P}(\lambda)$ avec $\lambda > 0$ inconnu. On a $R(\Omega) = \mathbb{N}$. Comme $\bigcup_{i=1}^{8} C_i = \{0, \dots, 7\} \neq \mathbb{N}$, on ajuste la dernière classe comme : $C_8 = \{7; 8; \dots; \infty\}$ (de sorte à ce que $\bigcup_{i=1}^{8} C_i = R(\Omega)$). Le paramètre λ étant inconnu, il faut l'estimer à l'aide des données. Comme $\mathbb{E}(X) = \lambda$, la méthode des moments nous assure qu'une estimation de λ est la moyenne x. On peut la calculer en faisant :

```
> nb = c(65, 110, 70, 48, 16, 5, 4, 2)
> lambda = sum((0:7) * nb) / 320
> lambda
[1] 1.628125
```

Cette estimation sera prise en compte dans la suite. Pour tout $i \in \{1; \dots; 7\}$, on a :

$$p_i = \mathbb{P}(R \in C_i) = \mathbb{P}(R = i - 1) = e^{-1.628125} \frac{1.628125^{i-1}}{(i-1)!}$$

De plus, on a :

$$p_8 = \mathbb{P}(R \in C_8) = \mathbb{P}(R \geq 7) = \sum_{i=7}^{\infty} e^{-1.628125} \frac{1.628125^{i}}{(i)!}$$

On peut obtenir ces probabilités avec les commandes :

```
> proba = c(dpois(0:6, lambda), 1 - ppois(6, lambda))
> proba
[1] 0.196297287 0.319596520 0.260171542 0.141197264 0.057471699 0.018714222
[7] 0.005078182 0.001473285
```

Dans un premier temps, on propose de mettre en oeuvre le test du Chi-deux en faisant :

```
> nb
[1]  65 110  70  48  16   5   4   2
> proba
[1] 0.196297287 0.319596520 0.260171542 0.141197264 0.057471699 0.018714222
[7] 0.005078182 0.001473285
> chisq.test(nb, p = proba)$p.value
[1] 0.1057037
Warning message:
In chisq.test(nb, p = proba) :
  l'approximation du Chi-2 est peut-être incorrecte
```

On dénombre alors deux problèmes :
— le logiciel n'a pas pris en compte le fait que λ a été estimé,
— il y a un "Warning message" nous avertissant que l'hypothèse : $np_i \geq 5$ pour tout $i \in \{1; \dots; 8\}$ n'est peut-être pas vérifiée.
Étudions ce dernier point :

```
> proba
[1] 0.196297287 0.319596520 0.260171542 0.141197264 0.057471699 0.018714222
[7] 0.005078182 0.001473285
> sum(nb)
[1] 320
> 320*prob
[1]  29.47981  70.29801  83.81686 136.40533
```

Comme $np_7 < 5$ et $np_8 < 5$, nous allons fusionner les classes C_6, C_7 et C_8, formant ainsi une nouvelle dernière classe : $C_6 = \{5; 6; \dots\}$ avec pour effectif : $n_6 = 5 + 4 + 2 = 11$. Il y a désormais $k = 6$ classes. On vérifie alors que l'hypothèse est vérifiée avec cette nouvelle configuration :

```
> nb2 = c(65, 110, 70, 48, 16, 11)
> proba2 = c(dpois(0:4, lambda), 1 - ppois(4, lambda))
> 320 * proba2
[1]  62.81513 102.27089  83.25489  45.18312  18.39094   8.08502
```

Aucune valeur ne dépasse 5. Pour avoir la p-valeur associée au test du Chi-deux en prenant en compte le fait que l'on a estimé un paramètre, on considère le degré de liberté :

$$\nu = k - 1 - l = 6 - 1 - 1 = 4$$

On fait :

```
> x2obs = chisq.test(nb2, p = proba2)$statistic
> deg = 4
> 1 - pchisq(x2obs, deg)
X-squared
0.3659448
```

On obtient alors la vraie p-valeur associée au test du Chi-deux (notons aussi qu'aucun "Warning message" n'apparaît).
Comme p-valeur > 0.05, les données ne nous permettent pas de rejeter H_0. Ainsi, on ne peut pas rejeter l'hypothèse selon laquelle X ne suit pas une loi de Poisson.

Exercice 6

Par l'énoncé, on observe la valeur de X sur chacun des n individus (appareils) d'un échantillon avec $n = 7$: $(x_1; : : : ; x_n)$ (avec $x_i \in \mathbb{R}$).
On considère les hypothèses :

— H_0 : "X suit la loi exponentielle $\varepsilon(0.01)$"

— H_1 : "X ne suit pas la loi exponentielle $\varepsilon(0.01)$".

On peut utiliser le test de Kolmogorov-Smirnov. On fait :

```
> x = c(145, 110, 170, 48, 116, 95, 74 )
> ks.test(x, "pexp", 0.01)$p.value
[1] 0.2005605
```

Comme p-valeur > 0.05, les données ne nous permettent pas de rejeter H_0. Ainsi, on ne peut pas rejeter l'hypothèse selon laquelle X suit la loi exponentielle $\varepsilon(0.01)$.

Exercice 7

Par l'énoncé, on observe la valeur de X sur chacun des n individus (attentes) d'un échantillon avec $n = 7$: $(x_1; \ldots ; x_n)$ (avec $x_i \in \mathbb{R}$). On détermine $1/\bar{x}$ en faisant :

```
> x = c(25.12, 12.36, 24.35, 12.19, 5.27, 18.35, 19.11, 27.08, 21.09, 17.19,
+       8.45, 13.27, 15.17)
> 1 / mean(x)
[1] 0.05936073
```

On considère les hypothèses :

— H_0 : "X suit la loi exponentielle $\varepsilon(0.05936073)$"

— H_1 : "X ne suit pas la loi exponentielle $\varepsilon(0.05936073)$"

On peut utiliser le test de Kolmogorov-Smirnov. On fait :

```
> ks.test(x, "pexp", 0.05936073)$p.value
[1] 0.05028391
```

Comme p-valeur > 0.05 (de justesse), les données ne nous permettent pas de rejeter H_0. Ainsi, on ne peut pas rejeter l'hypothèse selon laquelle X suit la loi exponentielle $\varepsilon(0.05936073)$.

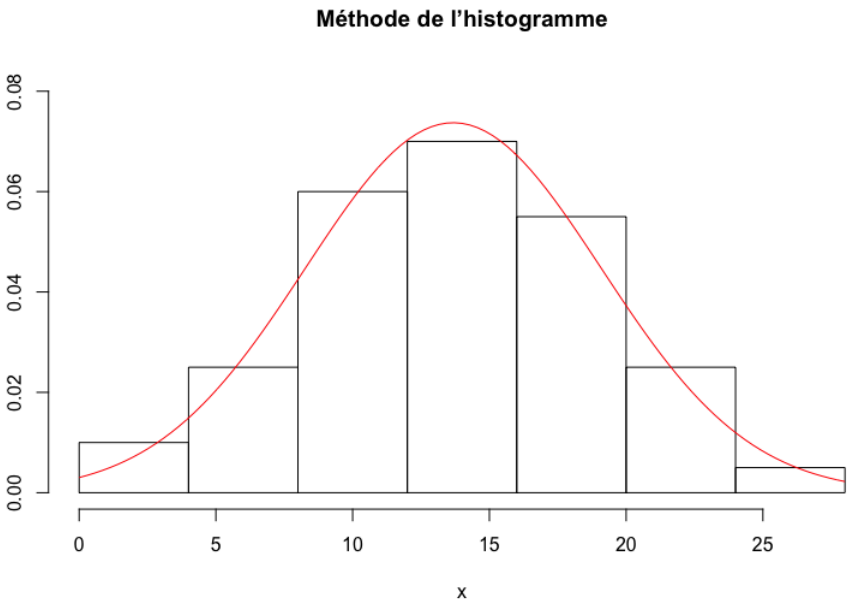
Exercice 8

Par l'énoncé, on observe la valeur de X sur chacun des n individus (pourceaux) d'un échantillon avec $n = 50$: $(x_1; \ldots; x_n)$ (avec $x_i \in \mathbb{R}$). Celles-ci sont présentées sous la forme d'un tableau classes-effectifs. On propose d'extrapoler les données en utilisant les centres des classes et les effectifs respectifs :

```
> xinf = c(0, 4, 8, 12, 16, 20, 24)
> xsup = c(4, 8, 12, 16, 20, 24, 28)
> centre = (xinf + xsup) / 2
> centre
[1]  2  6 10 14 18 22 26
> n = c(2, 5, 12, 14, 11, 5, 1)
> x = rep(centre, n)
> x
 [1]  2  2  6  6  6  6  6 10 10 10 10 10 10 10 10 10 10 10 10 14 14 14 14 14
[25] 14 14 14 14 14 14 14 14 14 18 18 18 18 18 18 18 18 18 18 18 22 22 22 22
[49] 22 26
```

Pour une analyse graphique, on peut utiliser la méthode de l'histogramme. On propose les commandes :

```
> hist(x, freq = FALSE, breaks = c(0, 4, 8, 12, 16, 20, 24, 28),
+      main = "Méthode de l'histogramme", ylim = c(0, 0.082), ylab = "")
> a = mean(x) ; b = sd(x)
> curve(dnorm(x, a, b), add = TRUE, col="red")
```

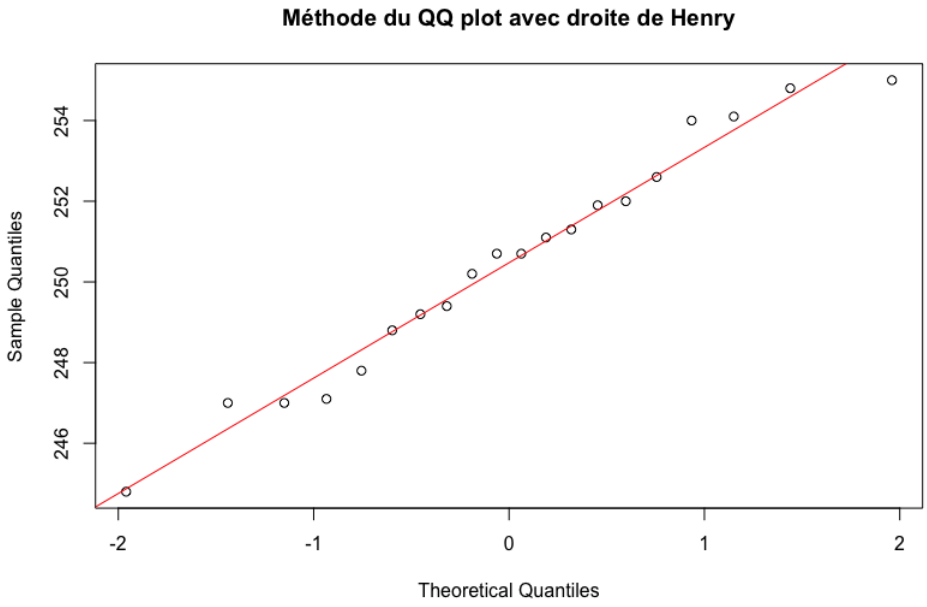


Les différences observées laissent penser que X suit une loi normale.

Exercice 9

Par l'énoncé, on observe la valeur de X sur chacun des n individus (plaquettes de beurre) d'un échantillon avec $n = 20$: $(x_1; \dots; x_n)$ (avec $x_i \in \mathbb{R}$).
Pour une analyse graphique, on peut utiliser la méthode du QQ plot avec la droite de Henry.
On propose les commandes :

```
x = c(247.0, 247.8, 250.2, 251.3, 251.9, 249.4, 248.8, 247.1, 255.0, 247.0,
254.8, 244.8, 250.7, 250.7, 252.6, 251.1, 254.1, 249.2, 252.0, 254.0)
qqnorm(x, main = "Méthode du QQ plot avec droite de Henry")
a = mean(x) ; b = sd(x)
curve(a + b * x, -6, 6, col = "red", add = TRUE)
```



On constate que la droite de henry ajuste bien le nuage de points, ce qui traduit le fait que X suit une loi normale. Confirmons cela avec un test statistique. On considère les hypothèses :

— H_0 : "X suit une loi normale"
— H_1 : "X ne suit pas une loi normale".

On peut utiliser le test de Shapiro-Wilk. On fait :

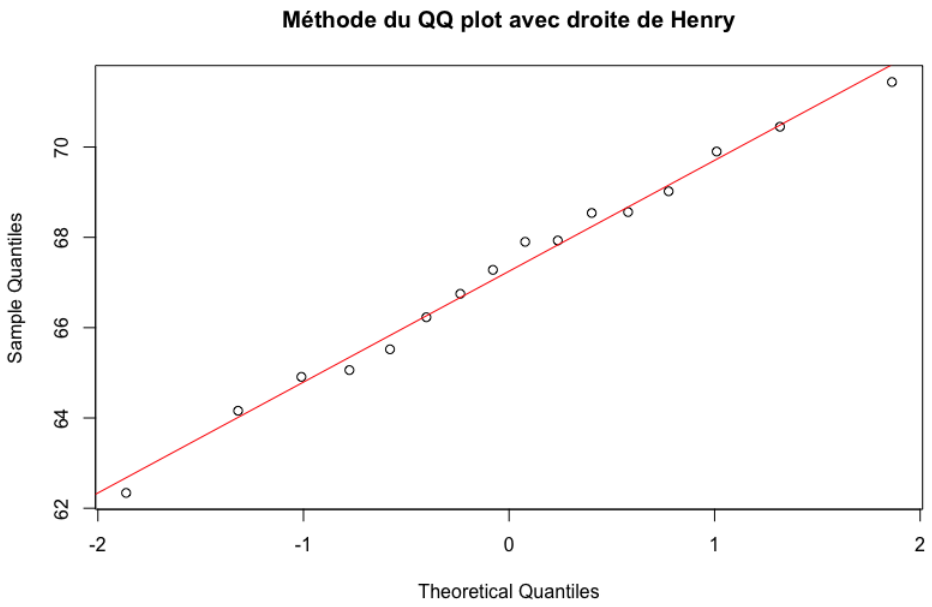
```
 [1] 247.0 247.8 250.2 251.3 251.9 249.4 248.8 247.1 255.0 247.0 254.8 244.8
[13] 250.7 250.7 252.6 251.1 254.1 249.2 252.0 254.0
> shapiro.test(x)$p.value
[1] 0.751598
```

Comme p-valeur > 0.05, on ne rejette pas H_0 ; l'hypothèse selon laquelle X suit une loi normale n'est pas rejetée.

Exercice 10

Par l'énoncé, on observe la valeur de X sur chacun des n individus (kiwis) d'un échantillon avec $n = 16$: $(x_1; \dots; x_n)$ (avec $x_i \in \mathbb{R}$).
Pour une analyse graphique, on peut utiliser la méthode du QQ plot avec la droite de Henry. On propose les commandes :

```
> x = c(65.06, 71.44, 67.93, 69.02, 67.28, 62.34, 66.23, 64.16, 68.56, 70.45,
+       64.91, 69.90, 65.52, 66.75, 68.54, 67.90)
> qqnorm(x, main = "Méthode du QQ plot avec droite de Henry")
> a = mean(x) ; b = sd(x)
> a;b
[1] 67.24938
[1] 2.455317
> curve(a + b * x, -6, 6, col = "red", add = TRUE)
```



On constate que la droite de henry ajuste bien le nuage de points, ce qui traduit le fait que X suit une loi normale. Confirmons cela avec un test statistique. On considère les hypothèses :

— H_0 : "X suit une loi normale"

— H_1 : "X ne suit pas une loi normale".

On peut utiliser le test de Shapiro-Wilk. On fait :

```
> x
 [1] 65.06 71.44 67.93 69.02 67.28 62.34 66.23 64.16 68.56 70.45 64.91 69.90
[13] 65.52 66.75 68.54 67.90
> shapiro.test(x)$p.value
[1] 0.9971045
```

Comme p-valeur > 0.05, on ne rejette pas H_0 ; l'hypothèse selon laquelle X suit une loi normale n'est pas rejetée.

Exercice 11

1. On propose :

```
> w = read.table("pression.txt", header = T)
> str(w)
'data.frame': 29 obs. of  2 variables:
 $ X1: int  39 45 47 65 46 67 42 67 56 64 ...
 $ Y : int  144 138 145 162 142 170 124 158 154 162 ...
> attach(w)
```

2. On considère les hypothèses :
   — H_0 : "Y suit une loi normale"
   — H_1 : "Y ne suit pas une loi normale"
   On fait le test de Shapiro-Wilk :

```
> shapiro.test(Y)$p.value
[1] 0.6421197
```

   Comme p-valeur > 0.05, l'hypothèse que Y suit une loi normale n'est pas rejetée.

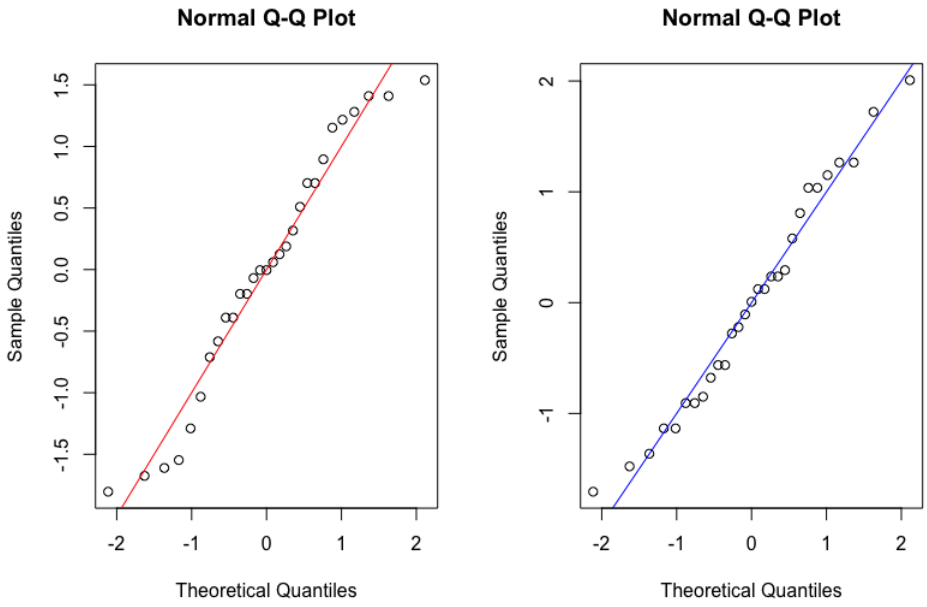
3. On considère les hypothèses :
   — H_0 : "X_1 suit une loi normale"
   — H_1 : "X_1 ne suit pas une loi normale".
   On fait le test de Shapiro-Wilk :

```
> shapiro.test(X1)$p.value
[1] 0.2083349
```

   Comme p-valeur > 0.05, l'hypothèse que X_1 suit une loi normale n'est pas rejetée.

4. On fait :

```
> par(mfrow = c(1, 2))
> qqnorm(scale(X1))
> abline(0, 1, col = "red")
> qqnorm(scale(Y))
> abline(0, 1, col = "blue")
```



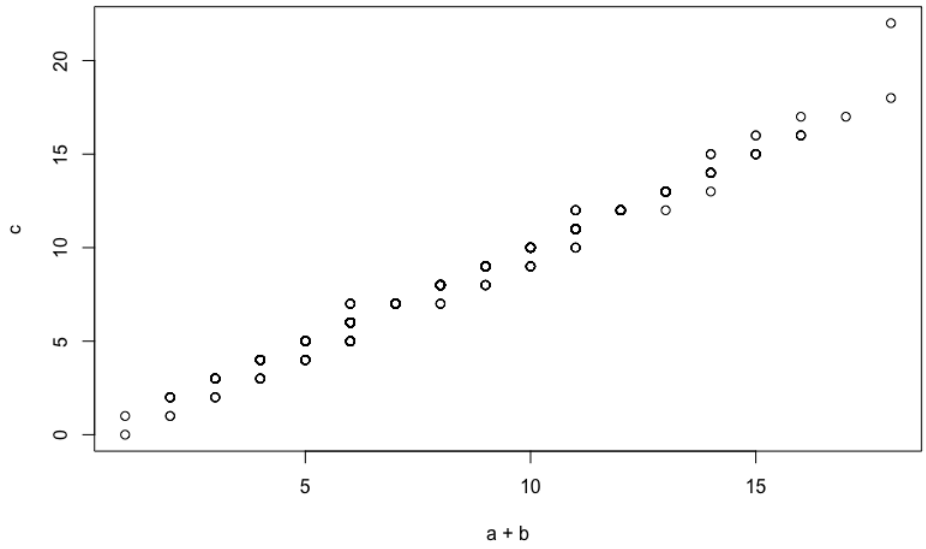
On analyse les QQ plot associés aux var X_1 et Y . On constate que les points sont presque alignés sur la droite diagonale d'équation $y = x$. On peut alors admettre que X_1 et Y suivent des lois normale.

Exercice 12

On fait :

```
> n = 1000
> a = rpois(n, 5)
> b = rpois(n, 3)
> c = rpois(n, 8)
> qqplot(a + b, c)
```

Cela renvoie :

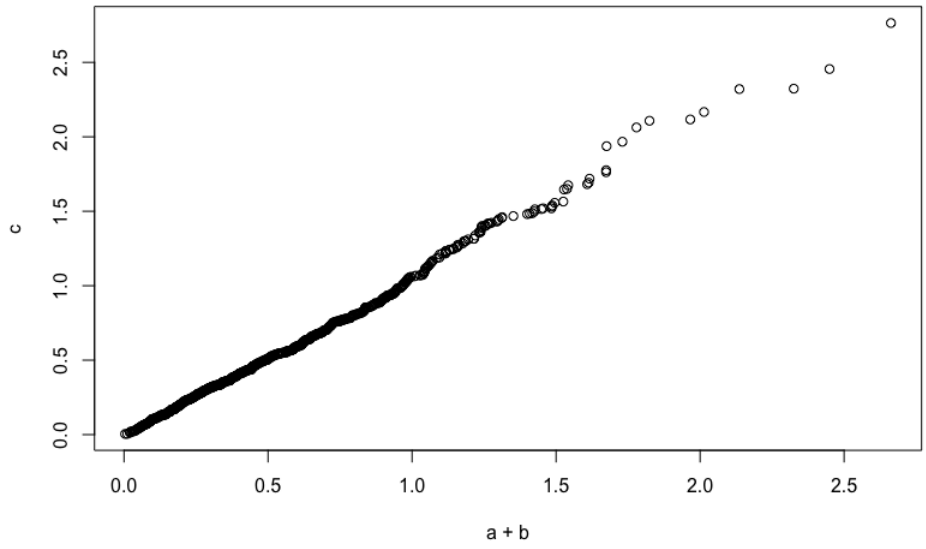


On constate que le nuage de points peut être ajusté par la droite $y = x$, ce qui illustre le résultat.

Exercice 13

— $X_1 \sim \varepsilon(\lambda)$,$X_2 \sim \varepsilon(\lambda)$, X_1 et X_2 indépendantes entraînent $X_1 + X_2 \sim \Gamma(\lambda, 2)$ avec $\lambda = 3.8$

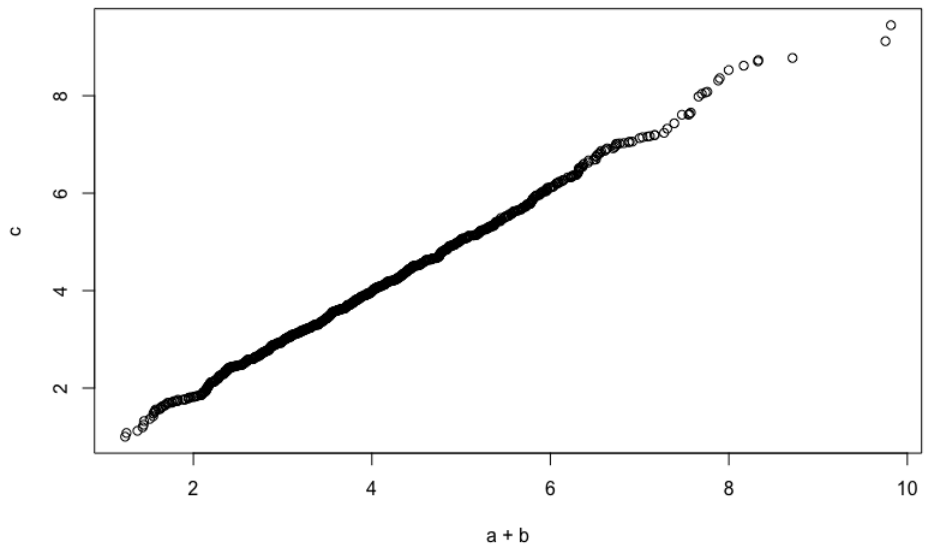
```
> a = rexp(1000, 3.8)
> b = rexp(1000, 3.8)
> c = rgamma(1000, 2, 3.8)
> qqplot(a + b, c)
```



On constate que le nuage de points peut être ajusté par la droite d'équation $y = x$, ce qui illustre le résultat.

— $X_1 \sim \Gamma(m_1, \lambda)$, $X_2 \sim \Gamma(m_2, \lambda)$, X_1 et X_2 indépendantes entraînent $X_1 + X_2 \sim \Gamma(m_1 + m_2, \lambda)$ avec $m_1 = m_2 = 4.2$ et $\lambda = 2.1$

```
> a = rgamma(1000, 4.2, 2.1)
> b = rgamma(1000, 4.2, 2.1)
> c = rgamma(1000, 8.4, 2.1)
> qqplot(a + b, c)
```

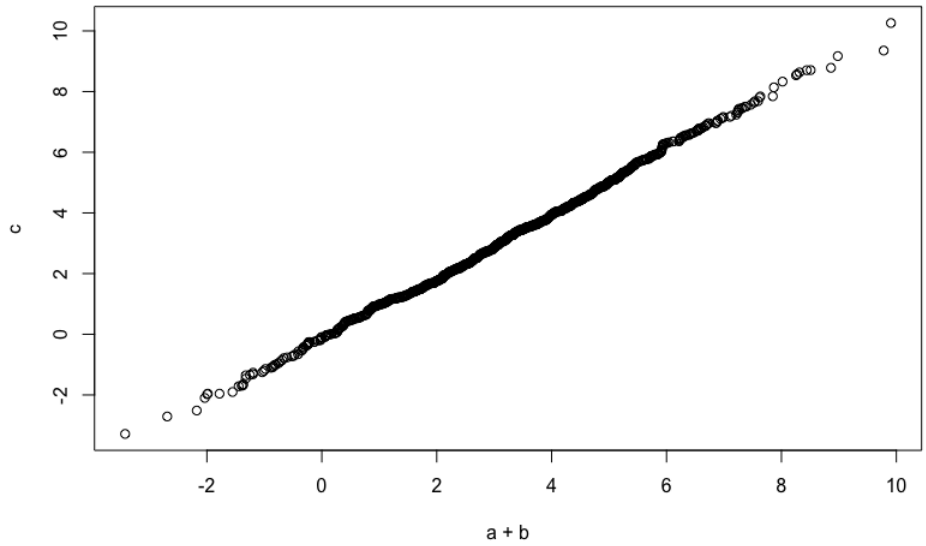


De nouveau, on constate que le nuage de points peut être ajusté par la droite d'équation $y = x$, ce qui illustre le résultat.

— $X_1 \sim (N)(\mu_1, \sigma_1^2)$,$X_2 \sim (N)(\mu_2, \sigma_2^2)$ X_1 et X_2 indépendantes entraînent $X_1 + X_2 \sim (N)(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$, avec $\mu_1 = \mu_2 = 1.6$ et $\sigma_1 = \sigma_2 = 1.5$

```
> a = rnorm(1000, 1.6, 1.5)
```

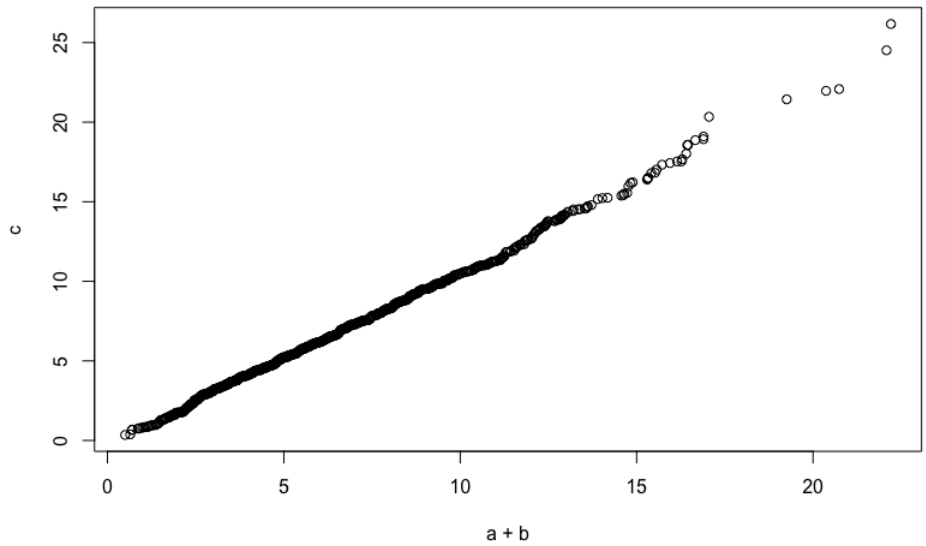
```
> b = rnorm(1000, 1.6, 1.5)
> c = rnorm(1000, 3.2, sqrt(1.5^2 + 1.5^2))
> qqplot(a + b, c)
```



De nouveau, on constate que le nuage de points peut être ajusté par la droite d'équation $y = x$, ce qui illustre le résultat.

— $X_1 \sim \chi^2(\nu_1)$,$X_2 \sim \chi^2(\nu_2)$, X_1 et X_2 indépendantes entraînent $X_1 + X_2 \sim \chi^2(\nu_1 + \nu_2)$ avec $\nu_1 = \nu_2 = 3.2$

```
> a = rchisq(1000, 3.2)
> b = rchisq(1000, 3.2)
> c = rchisq(1000, 6.4)
> qqplot(a + b, c)
```



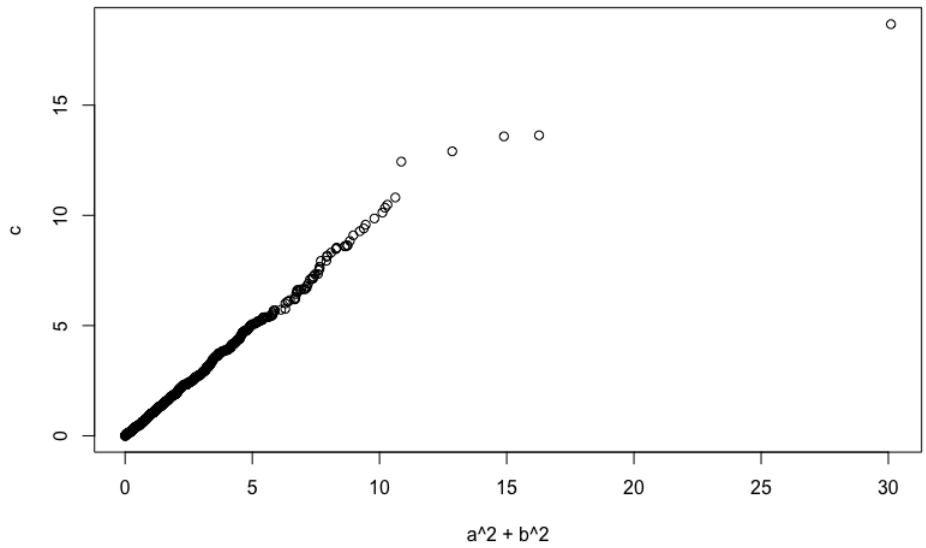
De nouveau, on constate que le nuage de points peut être ajusté par la droite d'équation y = x, ce qui illustre le résultat.

Exercice 14

— Caractérisation de la loi du chi-deux : $\chi^2(2)$:
Si $X \sim \mathcal{N}(0,1)$ et $Y \sim \mathcal{N}(0,1)$ alors

$$X^2 + Y^2 \sim \chi^2(2)$$

```
> a = rnorm(1000)
> b = rnorm(1000)
> c = rchisq(1000, 2)
> qqplot(a^2 + b^2, c)
```



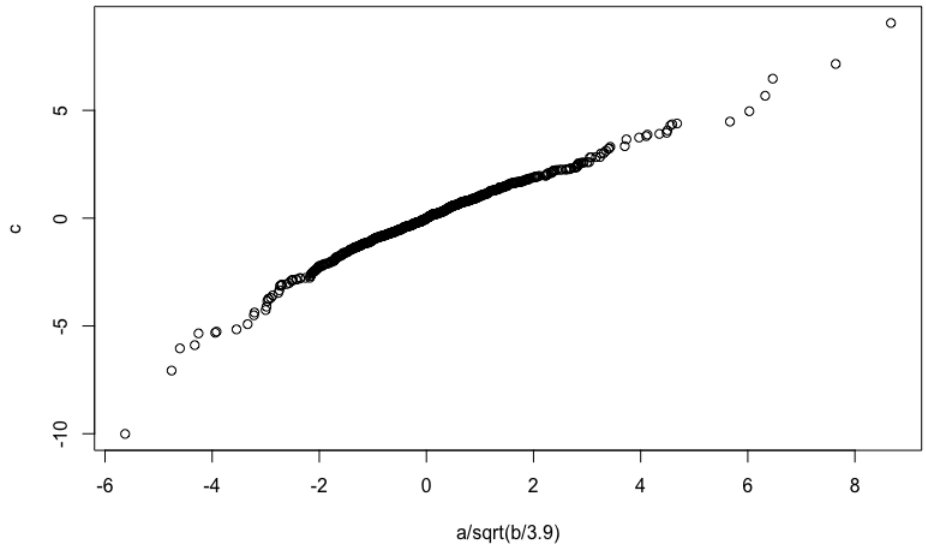
On constate que le nuage de points peut être ajusté par la droite d'équation $y = x$, ce qui illustre le résultat.

— Caractérisation de la loi de Student $\mathcal{T}(\nu)$:
Si $X \sim \mathcal{N}(0,1)$ et $Y \sim \chi^2(\mu)$ alors

$$\frac{X}{\sqrt{\frac{Y}{\nu}}} \sim \mathcal{T}(\nu)$$

avec $\nu = 3.9$

```
> a = rnorm(1000)
> b = rchisq(1000, 3.9)
> c = rt(1000, 3.9)
> qqplot(a / sqrt(b / 3.9), c)
```



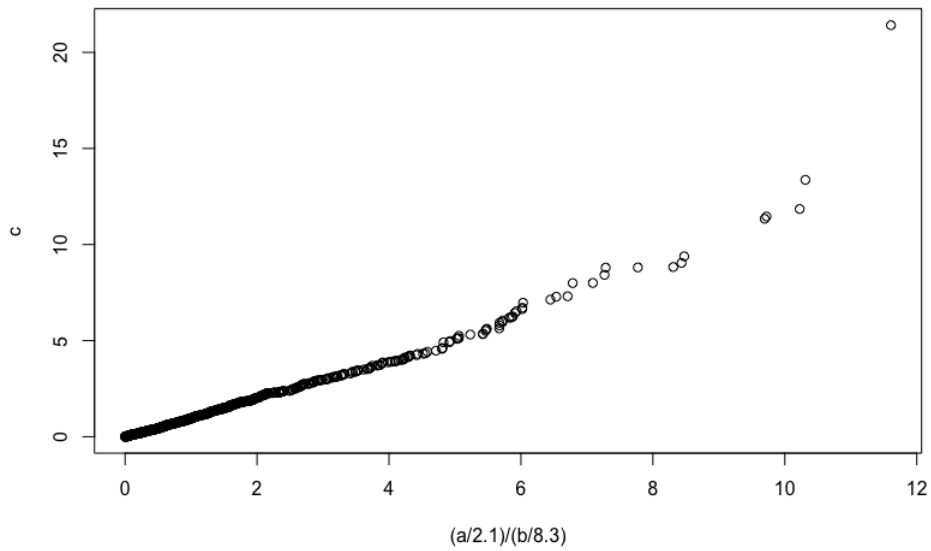
De nouveau, on constate que le nuage de points peut être ajusté par la droite d'équation $y = x$, ce qui illustre le résultat.

— Caractérisation de la loi de Fisher $\mathcal{F}(\nu_1, \nu_2)$

$$\frac{\frac{X}{\nu_1}}{\frac{Y}{\nu_2}} = \frac{\nu_2 X}{\nu_1 Y} \sim \mathcal{F}(\nu_1, \nu_2)$$

avec $(\nu_1, \nu_2) = (2,1; 8,3)$.

```
> a = rchisq(1000, 2.1)
> b = rchisq(1000, 8.3)
> c = rf(1000, 2.1, 8.3)
> qqplot((a / 2.1) / (b / 8.3), c)
```



De nouveau, on constate que le nuage de points peut être ajusté par la droite d'équation $y = x$, ce qui illustre le résultat.

Exercice 15

Dans les commandes, on génère 100 observations d'une var X suivant la loi normale centrée réduite $\mathcal{N}(0;1)$. On les place dans un vecteur x.
Ensuite, on créé un vecteur numérique a. Dans la boucle for, pour tout $i \in \{1;\dots;6\}$, on considère les hypothèses suivantes :

— $H_0(i)$: "X suit la loi normale $\mathcal{N}(0,(1+\frac{i-1}{10})^2)$"

— $H_1(i)$"X ne suit pas la loi normale $\mathcal{N}(0,(1+\frac{i-1}{10})^2)$"

on utilise le test de Kolmogorov-Smirnov, on calcule la p-valeur associée à $H_0(i)$ et on met cette valeur au i-ème élément de a.
Ensuite, on affiche a.
On remarque alors que, logiquement, plus i est grand, plus $1+(i-1)/10$ s'éloigne de 1, plus la p-valeur est petite, moins on a de certitude en affirmant que la loi de X est en adéquation avec la loi normale $\mathcal{N}(0,(1+\frac{i-1}{10})^2)$

IV Div'R

V Représentations graphiques

VI Programmation avec R

VII Fonctions usuelles et aide mémoire

2 Lois

Loi Normale Centrée Réduite $\mathcal{N}(0;\,1)$.

$$\mathbb{P}(t) = P(X \leq t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}} dx \text{ et } \mathbb{P}(-t) = 1 - \mathbb{P}(t).$$

2.1 Loi Normale

t	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
q 3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

2.2 Loi Normale

3 R datasets informations

```
> help(nomDuDataset)
```

— *"ToothGrowth"* : The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, (orange juice or ascorbic acid (a form of vitamin C and coded as VC).

1. len numeric Tooth length
2. supp factor Supplement type (VC or OJ).
3. dose numeric Dose in milligrams/day

— *"economics"* : This dataset was produced from US economic time series data available from research.stlouisfed.org. economics is in "wide" format, economics_long is in "long" format.

1. date : Month of data collection
2. psavert : personal savings rate
3. pce : personal consumption expenditures, in billions of dollars
4. unemploy : number of unemployed in thousands
5. uempmed : median duration of unemployment, in weeks
6. pop :total population, in thousands

— *"diamonds"* : A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows :

1. price : price in US dollars ($326 ?$18,823)
2. carat : weight of the diamond (0.2 ?5.01)
3. cut : quality of the cut (Fair, Good, Very Good, Premium, Ideal)
4. color : diamond colour, from J (worst) to D (best)
5. clarity : a measurement of how clear the diamond is (I1 (worst), SI1, SI2, VS1, VS2, VVS1, VVS2, IF (best))
6. x : length in mm (0 ?10.74)
7. y : width in mm (0 ?58.9)
8. z : depth in mm (0 ?31.8)
9. depth : total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43 ?79)
10. table : width of top of diamond relative to widest point (43 ?95)

— *"world2"* : This is an alternative version of the world database based on latitudes [0, 360), which then has the Pacific Ocean in the centre of the map.
The data file is merely a character string which specifies the name of an environment variable which contains the base location of the binary files used by the map drawing functions. This environment variable (R_MAP_DATA_DIR_WORLD for the datasets in the maps package) is set at package load time if it does not already exist. Hence setting the environment variable before loading the package can override the default location of the binary datasets.

— *lung* : Survival in patients with advanced lung cancer from the North Central Cancer Treatment Group. Performance scores rate how well the patient can perform usual daily activities.

1. inst : Institution code
2. time : Survival time in days
3. status : censoring status 1=censored, 2=dead
4. age : Age in years
5. sex : Male=1 Female=2
6. ph.ecog : ECOG performance score (0=good 5=dead)
7. ph.karno : Karnofsky performance score (bad=0-good=100) rated by physician
8. pat.karno : Karnofsky performance score as rated by patient
9. meal.cal : Calories consumed at meals
10. wt.loss : Weight loss in last six months

— *mtcars* : The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973 ?74 models).

1. mpg Miles/(US) gallon
2. cyl Number of cylinders
3. disp Displacement (cu.in.)
4. hp Gross horsepower
5. drat Rear axle ratio
6. wt Weight (1000 lbs)
7. qsec 1/4 mile time
8. vs V/S
9. am Transmission (0 = automatic, 1 = manual)
10. gear Number of forward gears
11. carb Number of carburetors

— *mpg :* This dataset contains a subset of the fuel economy data that the EPA makes available on fueleconomy.gov. It contains only models which had a new release every year between 1999 and 2008 - this was used as a proxy for the popularity of the car.

1. manufacturer : manufacturer name
2. model : model name
3. displ : engine displacement, in litres
4. year : year of manufacture
5. cyl : number of cylinders
6. trans : type of transmission
7. drv : f = front-wheel drive, r = rear wheel drive, 4 = 4wd
8. cty : city miles per gallon
9. hwy : highway miles per gallon
10. fl : fuel type
11. class : "type" of car

— *iris* : This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica. iris is a data frame with 150 cases (rows) and 5 variables (columns) named Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species.
iris3 gives the same data arranged as a 3-dimensional array of size 50 by 4 by 3, as represented by S-PLUS. The first dimension gives the case number within the species subsample, the second the measurements with names Sepal L., Sepal W., Petal L., and Petal W., and the third the species.

— *faithful :* Waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.

1. eruptions numeric Eruption time in mins
2. waiting numeric Waiting time to next eruption (in mins)