Équipe no 4

Nicholas Langevin
(111 184 631)

Alexandre Turcotte
(111 172 613)

Mathématiques actuarielles IARD 1
ACT-2005

Travail pratique 1

Travail présenté à
Andrew Luong

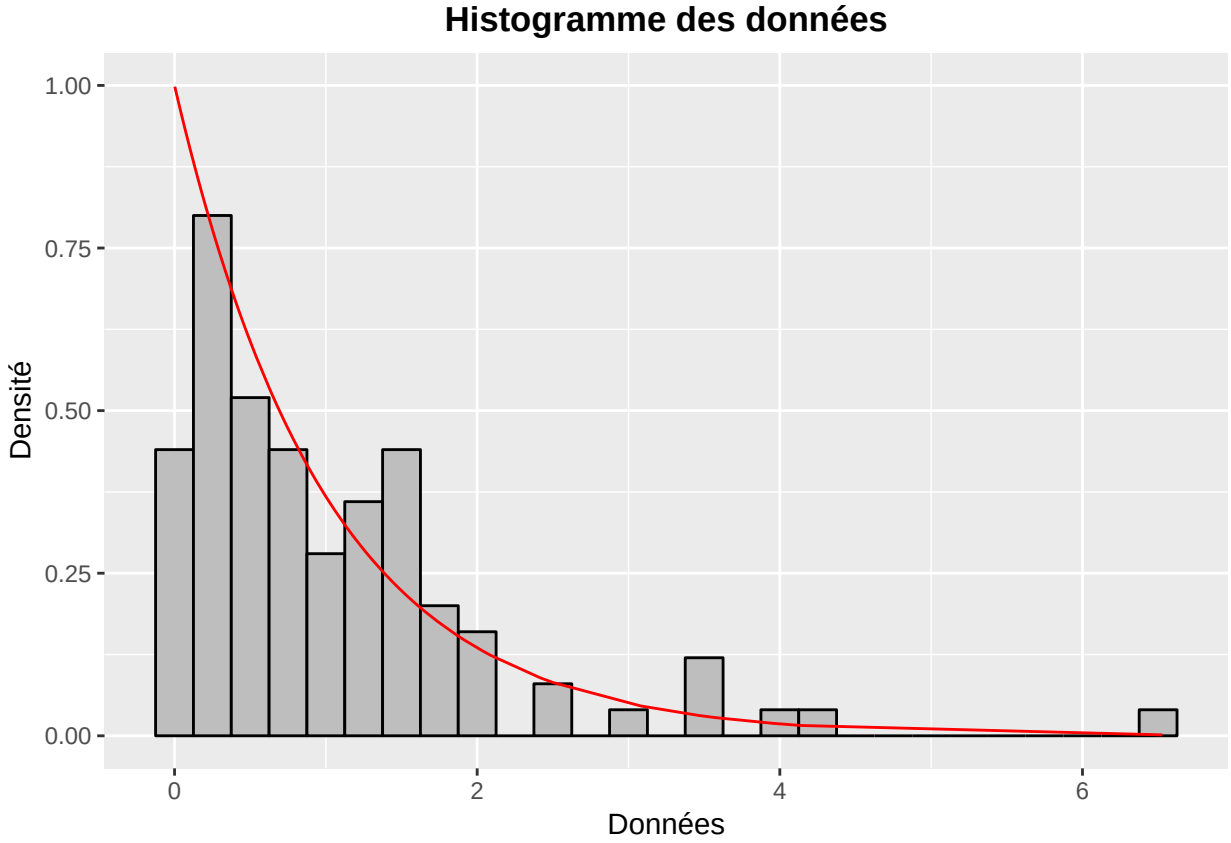
École d'actuariat
Université Laval
Automne 2018

Question 1

a) Estimation du coefficient d'asymétrie

100 données ont été simulées à l'aide d'une loi exponentielle de moyenne 1 et ces données se trouvent dans l'annexe. D'ailleurs, à partir de ces données, il a été possible de tracer un histrogramme et d'estimer le coefficient d'asymétrie. La densité de la loi exponentielle de moyenne 1 a été ajouté à l'histogramme afin de pouvoir comparer la distribution des données.

$$\hat{\gamma} = \frac{\sum_{i=1}^{100}(X_i - \mu)^3}{\sigma^3} \;=\; 2.1802922$$



À partir de l'histogramme, il est possible de noter que la distribution est plus dense et concentrée à gauche et que la queue de la distribution tend vers la droite. Par conséquent, la distribution n'est pas symétrique et le coefficient d'asymétrie devrait être positif. Cela concorde effectivement avec le coefficient d'asymétrie estimé empiriquement puisqu'il est de 2.1802922 conparativement à celui de la loi normale qui est égal à 0. La loi normale a une distribution symétrique et comme le coefficient d'asymétrie des valeurs simulées est plus grand que celui de la loi normale, cela implique que la distribution est asymétrique vers la droite. Donc, elle possède une queue de distribution à droite comme il est possible d'observer sur l'histogramme précédent.

b) Intervalle de confiance pour le coefficient d'asymétrie

Avec la méthode de ré-échantillonnage, la variance estimée pour le coefficient d'asymétrie est de 0.1144019. À partir de l'estimateur ponctuel du coefficient d'asymétrie calculé en a) et de cette variance estimée, il est possible d'obtenir l'intervalle de confiance suivant pour le coefficient d'asymétrie :

$$[1.5173668, 2.8432177]$$

c) Coefficient d'asymétrie théorique

Les moments de la loi exponentielle sont donnés par:

$$\begin{aligned} E[x] &= M_x^{'}(t)\Big|_{t=0} \\ &= \frac{d}{dt}\left(\frac{\theta}{\theta - t}\right)\Big|_{t=0} \\ &= \left(\frac{\theta}{(\theta - t)^2}\right)\Big|_{t=0} \\ &= \frac{1}{\theta} \end{aligned} \qquad \begin{aligned} E[x^2] &= M_x^{''}(t)\Big|_{t=0} \\ &= \frac{d}{dt}\left(\frac{\theta}{(\theta - t)^2}\right)\Big|_{t=0} \\ &= \left(\frac{2\theta}{(\theta - t)^3}\right)\Big|_{t=0} \\ &= \frac{2}{\theta^2} \end{aligned} \qquad \begin{aligned} E[x^3] &= M_x^{'''}(t)\Big|_{t=0} \\ &= \frac{d}{dt}\left(\frac{2\theta}{(\theta - t)^3}\right)\Big|_{t=0} \\ &= \left(\frac{6\theta}{(\theta - t)^4}\right)\Big|_{t=0} \\ &= \frac{6}{\theta^3} \end{aligned}$$

Ainsi, le coeficient d'asymétrie théorique est donnée par

$$\begin{aligned} \gamma &= \frac{E\left[(x-\mu)^3\right]}{\sigma^3} \\ &= \frac{1}{\sigma^3}\left(E[x^3 - 3x^2\mu + 3x\mu^2 - \mu^3]\right) \\ &= \frac{1}{\sigma^3}\left(E[x^3] - 3\mu E[x^2] + 3\mu^2 E[x] - \mu^3\right) \\ &= \theta^3\left[\frac{6}{\theta^3} - 3\left(\frac{1}{\theta}\right)\left(\frac{2}{\theta^2}\right) + 3\left(\frac{1}{\theta}\right)^3 - \left(\frac{1}{\theta}\right)^3\right] \\ &= \theta^3\left[\frac{6}{\theta^3} - \frac{6}{\theta^3} + \frac{2}{\theta^3}\right] \\ &= 6 - 6 + 2 \\ &= 2 \end{aligned}$$

Le coefficient d'asymétrie théorique de la loi exponentielle de moyenne 1 est de 2. Cela s'avère compatible avec l'estimé ponctuel obtenu en a) puisque celui-ci est de 2.1802922, ce qui est assez près de 2. Pour ce qui est de l'intervalle de confiance obtenu en b), il est possible d'observer que la valeur théorique est incluse dans cette intervalle qui est (1.5173668, 2.8432177). Par conséquent, les estimateurs obtenus sont compatibles avec la valeur théorique. Toutefois, ceux-ci ne sont pas très précis en raison du nombre d'observations qui est assez faible, soit 100 observations. En augmentant le nombre d'observation, l'intervalle de confiance serait plus petit et l'estimateur serait plus précis.

Question 2

a) Estimation de l'espérance limitée

Pour pouvoir déterminer les valeurs des limites u, il faut utiliser la fonction quantile théorique d'une loi exponentielle ($\theta = 1$). Ainsi, en ayant u, il sera possible d'estimer l'espérance limitée.

$$\begin{aligned} X &\sim Exp(\theta = 1) \\ F(x) &= 1 - e^{-\frac{x}{\theta}} \\ &= 1 - e^{-x} \\ k &= 1 - e^{-x} \\ 1 - k &= e^{-x} \\ F_x^{-1}(k) &= x = -ln(1-k) \\ \hat{E}\left[min(X,u)\right] &= \frac{\sum_{i=1}^{100} min(X,\mu)}{n} \end{aligned}$$

	Percentile	Limite u	Espérance limitée
1	0.25	0.2876821	0.2514125
2	0.35	0.4307829	0.3528583
3	0.50	0.6931472	0.5030752
4	0.60	0.9162907	0.6086610
5	0.75	1.3862944	0.7788036
6	0.85	1.8971200	0.8728915

Table 1: Valeurs de l'espérance limitée pour chaque limite u donnée

$E[min(X,u)]$ estimé est une fonction croissante en fonction du u. En effet, le tableau présente une augmentation de la valeur de l'espérance lorsque u augmente. Cela s'avère tout à fait logique puisque lorsque u est petit, la fonction $min(X,u)$ prend davantage en considération les valeurs de u. Par conséquent, les valeurs supérieures de x sont réduites, ce qui réduit donc la moyenne, car elle ne tient compte que des valeurs inférieures ou égales à u. Alors, si u augmente, l'espérance prendra en compte des valeurs plus grande de x, car u aura augmenté.

b) Intervalle de confiance pour l'espérance limitée

	Percentile	Limite u	Espérance limité	Variance
1	0.50	0.6931472	0.5030752	0.00058433
2	0.75	1.3862944	0.7788036	0.00238501

Table 2: Valeurs de l'espérance limitée et de sa variance pour les percentiles 0.5 et 0.75

Avec la méthode de ré-échantillonnage, la variance estimée pour $E[min(X, F^{-1}(0.5))]$ est de 0.00058433, alors que celle pour $E[min(X, F^{-1}(0.75))]$ est de 0.002385. À partir des estimateurs de $E[min(X,u)]$ calculés en a) pour $u = F^{-1}(0.5)$ et $u = F^{-1}(0.75)$ et de leurs variances estimées, il est possible d'obtenir les intervalles de confiance suivants :

$$[0.4556972, 0.5504532],\ pour\ u = F^{-1}(0.5)$$

$$[0.6830857, 0.8745215],\ pour\ u = F^{-1}(0.75)$$

c) Espérance limitée théorique

$$
\begin{aligned}
X &\sim Exp(\theta = 1) \\
S(x) &= e^{-x}, \text{ où } x > 0 \\
u &= F_x^{-1}(k) = -ln(1-k) \\
E\left[min(X,u)\right] &= E\left[X \wedge u\right] \\
&= \int_0^u x f_x(x)dx + \int_0^u u f_x(x)dx \\
&= \int_0^u S_x(x)dx \\
&= \int_0^u e^{-x}dx \\
&= [-e^{-x}]_0^u \\
&= 1 - e^{-u} \\
&= 1 - e^{-(-ln(1-k))} \\
&= 1 - (1-k) \\
&= k
\end{aligned}
$$

Ainsi, les valeurs théoriques des espérances limitées sont :

$$
\begin{aligned}
E\left[min(X, F_x^{-1}(0.5))\right] &= 0.5 \\
E\left[min(X, F_x^{-1}(0.75))\right] &= 0.75
\end{aligned}
$$

L'espérance limitée théorique de la loi exponentielle de moyenne 1 est égale à son quantile. Par conséquent, $E[min(X, F^{-1}(0.5))] = 0.5$ et $E[min(X, F^{-1}(0.75))] = 0.75$. Cela s'avère compatible avec les valeurs estimées obtenues en a) puisque ceux-ci sont de 0.5030752 et 0.7788036, ce qui est assez près des valeurs théoriques. Pour ce qui est des intervalles de confiance obtenu en b), il est possible d'observer que les valeurs théoriques sont incluses dans chacun des intervalles respectifs qui sont [0.4556972, 0.5504532], pour le premier, et [0.6830857, 0.8745215], pour le second. Par conséquent, les estimateurs obtenus sont compatibles avec les valeurs théoriques. Il est donc possible d'affirmer que ces estimateurs non-paramétriques sont assez performants puisque les valeurs obtenues sont très près de la valeur théorique.

Question 3

a) Détermination de la fonction de survie à l'aide de l'estimateur Kaplan-Meier

Valeurs estimées de survie de Kaplan-Meier						
temps		Survie	Echec	Erreur type de survie	Nombre d'échecs	Nombre restant
0.000		1.0000	0	0	0	10
30.000		0.9000	0.1000	0.0949	1	9
40.000		0.8000	0.2000	0.1265	2	8
57.000		0.7000	0.3000	0.1449	3	7
65.000		0.6000	0.4000	0.1549	4	6
65.000	*	.	.	.	4	5
84.000		0.4800	0.5200	0.1640	5	4
90.000		0.3600	0.6400	0.1610	6	3
92.000	*	.	.	.	6	2
98.000		0.1800	0.8200	0.1506	7	1
101.000		0	1.0000	.	8	0

Figure 1: Données permettant de trouver les estimateurs Kaplan-Meier, et ce, à partir de SAS

La table précèdente présente toutes les données utiles permettant de déterminer l'estimateur Kaplan-Meier. Par conséquent, il est possible de calculer cet estimateur à partir de sa définition :

$$\hat{S}_n(t) = \begin{cases} 1, \ 0 \leqslant t < y, \\ \prod_{i=1}^{j-1} (\frac{r_i - S_i}{r_i}), \ y_{j-1} \leqslant t < y_i, \\ \prod_{i=1}^{k} (\frac{r_i - S_i}{r_i}), \ t > y_k \end{cases}$$

$$\hat{S}_n(t) = \begin{cases} 1 & , 0 \leqslant t < 30, \\ 0.9 & , 30 \leqslant t < 40, \\ 0.8 & , 40 \leqslant t < 57, \\ 0.7 & , 57 \leqslant t < 65, \\ 0.6 & , 65 \leqslant t < 84, \\ 0.48 & , 84 \leqslant t < 90, \\ 0.36 & , 90 \leqslant t < 98, \\ 0.18 & , 98 \leqslant t < 101, \\ 0 & , t \geqslant 101 \end{cases}$$

b) Graphique de l'estimateur Kaplan-Meier et intervalle de confiance pour $\hat{S}_n(50)$

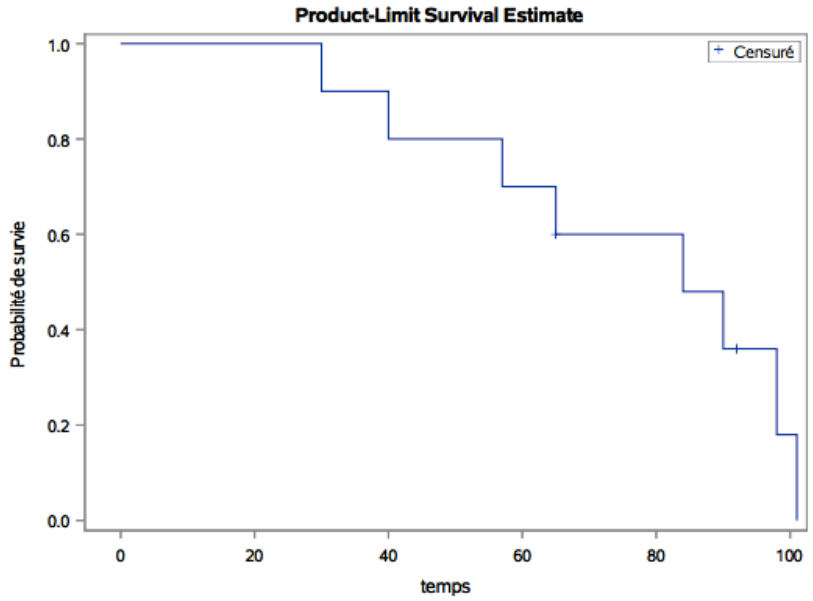


Figure 2: Valeur de la fonction de survie estimée à l'aide de l'estimateur Kaplan-Meier en fonction des temps de décès

Calculer un estimateur pour la variance de la fonction de survie à l'aide de la formule de Greenwood :

$$\widehat{Var}(\hat{S}_n(t)) = [\hat{S}_n(t)]^2 \sum_{i=1}^{j} \frac{S_i}{r_i(S_i - r_i)}$$

$$\widehat{Var}(\hat{S}_n(t)) = \begin{cases} 0 & ,0 \leqslant t < 30, \\ 0.009 & ,30 \leqslant t < 40, \\ 0.016 & ,40 \leqslant t < 57, \\ 0.021 & ,57 \leqslant t < 65, \\ 0.024 & ,65 \leqslant t < 84, \\ 0.02688 & ,84 \leqslant t < 90, \\ 0.02592 & ,90 \leqslant t < 98, \\ 0.02268 & ,98 \leqslant t < 101 \end{cases}$$

Alors, en ayant les valeurs de l'estimateur Kaplan-Meier et celles de Greenwood, il est possible de déduire les intervalles de confiance au niveau 0.95 pour $\hat{S}_n(50)$. Donc, la valeur de $\hat{S}_n(50) = 0.8$ et celle de la $\widehat{Var}(\hat{S}_n(50)) = 0.016$ puisque 50 appartient à l'intervalle 40 à 57 et puisqu'il s'agit d'une fonction de survie, il faut utiliser la borne inférieure. La figure 2 montre également que 50 est sur la marche correspondant à 0.8.

$$\begin{aligned} &S_n(50) \in \hat{S}_n(50)) \pm z_{0.975}\sqrt{\widehat{Var}(\hat{S}_n(50))} \\ &S_n(50) \in 0.8 \pm z_{0.975}\sqrt{0.016} \\ &S_n(50) \in [0.552082, 1.047918] \end{aligned}$$

c) Intervalle de confiance pour $\hat{S}_n(50)$ avec la transformation log (-log)

La valeur estimée de la fonction de survie et la valeur estimée de la variance sont toujours les mêmes que celles de la section b). Par conséquent, seulement l'intervalle de confiance est modifié. Alors, l'intervalle de confiance log transformed au niveau de confiance 95% est déterminé de cette façon :

$$\hat{S}_n(50) \in [\hat{S}_n(50))^{1/u}, \hat{S}_n(50))^u]$$

$$o\grave{u}\ u = exp\left\{\frac{z_{0.975}\sqrt{\widehat{Var}[\hat{S}_n(50)]}}{\hat{S}_n(50)\ln(\hat{S}_n(50))}\right\}$$

$$= exp\left\{\frac{z_{0.975}\sqrt{0.016}}{0.8 \cdot \ln(0.8)}\right\}$$

$$\hat{S}_n(50) \in [0.4086908, 0.9458726]$$

Les deux intervalles de confiance contiennent la valeur estimée de la fonction de survie. D'ailleurs, les deux intervalles de confiance sont assez larges, soit un étendu de 0.495836 pour l'intervalle de confiance normal et de 0.5371819 pour l'intervalle de confiance avec la transformation log. Alors, le second intervalle est légèrement plus grand et cela s'explique par le fait que le nombre d'observations est plutôt faible. En ayant un nombre d'observations plus élevé, l'étendu des intervalles de confiance seraient plus faibles. D'ailleurs, l'intervalle de confiance avec la transformation log serait plus précis que celui normal, car la convergence vers la loi normale se fait plus rapidement. Finalement, l'intervalle de confiance avec la transformation log donne une borne supérieure qui a plus de sens, car elle est inférieure à 1 et la valeur d'une fonction de survie doit se trouver entre 0 et 1.

Annexe

Question 1

a) Estimation du coefficient d'asymétrie

Échantillon de 100 données simulées à partir d'une loi exponentielle ($\theta = 1$) :

```
data <-  rexp(100, 1)
```

```
##   [1] 0.488657233 0.339177388 0.736506375 1.438225010 0.174769304
##   [6] 0.070925108 1.057122434 0.788101510 1.112217082 0.181316652
##  [11] 3.544854766 2.090703877 0.047749216 0.494062895 4.142850792
##  [16] 0.860649363 1.406497737 0.045046159 1.397286030 0.660906192
##  [21] 0.272923105 0.371098046 3.961818268 0.469019631 0.852715764
##  [26] 3.466884365 0.995280062 0.222400230 0.459857417 1.469787422
##  [31] 1.535016322 0.343007872 1.135763040 1.132503868 0.289504916
##  [36] 0.137540061 0.246590225 0.010301427 1.588164035 0.358647224
##  [41] 1.891250648 0.425798224 0.102364165 0.278363570 0.003788391
##  [46] 1.653295521 2.503299942 0.458441574 0.652582877 1.401962474
##  [51] 1.855782070 0.678851549 0.169078148 0.383953454 1.172795020
##  [56] 1.322907019 1.199353934 0.016570793 0.097131792 1.738748383
##  [61] 2.036200609 0.442668886 2.415775323 1.870185513 0.698428379
##  [66] 1.290145343 1.598015681 0.907234591 0.125813944 0.367951234
##  [71] 0.028502554 1.128864091 0.505788688 0.001886764 1.385176676
##  [76] 3.583714757 1.119854477 1.450977566 1.646787436 0.261256020
##  [81] 1.011040086 0.544445592 0.417613582 3.081648424 0.205689473
##  [86] 1.243885942 0.941053231 0.650554588 0.295850069 0.748657196
##  [91] 6.527928812 0.326071966 0.024820297 0.264905422 0.625615774
##  [96] 1.517615159 1.266849371 0.376533218 1.916030822 0.540617309
```

```
# Fonction pour calculer le coefficient d'asymétrie.
coef_asymetrie <- function(x){
    mu <- mean(x)
    sd <- sd(x)
    mean((x - mu)^3) / sd^3
}

# Estimation du coefficient d'asymétrie des données.
estimateur_coef_asymetrie <- coef_asymetrie(data)

# Histograme des données avec la courbe théorique d'un loi exponentielle (teta = 1)
library(ggplot2)
data2 <- data.frame(data)
df <- data.frame(x = data, y = dexp(data, 1))

ggplot(data = data2) +
        geom_histogram(aes(x = data, y = ..density..),
                       binwidth = 0.25, fill = "grey", color = "black") +
        geom_line(data = df, aes(x = data, y = y), color = "red") +
        ggtitle("Histogramme des données") +
        theme(plot.title =  element_text(face="bold", hjust = 0.5)) +
        xlab("Données") + ylab("Densité") +
        scale_y_continuous(limits = c(0, 1))
```

b) Intervalle de confiance pour le coefficient d'asymétrie

```r
# Estimation du theta
teta <- mean(data)

# Simulation de 50 échantillions
echantillion <- lapply(1:50, function(i) rexp(100, teta))

# Estimation des 50 coefficients d'asymétrie
coef_asymetrie_simul <- sapply(1:50, function(i) coef_asymetrie(echantillion[[i]]) )

# Estimation de la variance empirique
variance_coef_asymetrie <- var(coef_asymetrie_simul)

# Interval de confiance pour le coefficient d'asymétrie
IC_coef_asymetrie <- cbind(estimateur_coef_asymetrie -
                               qnorm(0.975) * sqrt(variance_coef_asymetrie),
                           estimateur_coef_asymetrie +
                               qnorm(0.975) * sqrt(variance_coef_asymetrie))
```

c) Coefficient d'asymétrie théorique

Aucun calcul R n'a été fait dans cette section.

Question 2

a) Estimation de $E[min(X,u)]$

```
# Déterminer les limites u à l'aide de la fonction quantile
k <- c(0.25, 0.35, 0.5, 0.6, 0.75, 0.85)
fonction_quantile <- function(x) -log(1-x)
limite_u <- sapply(k, function(i) fonction_quantile(i))

# Déterminer l'espérance limitée pour chacune des limites u
Estimateur_Esperance_limite <- sapply(1:length(limite_u), function(u)
                                    mean(sapply(1:100, function(i)
                                        min(data[i], limite_u[u]))))

# Publier les résultats pour chaque u
resultats <- data.frame(k, limite_u, Estimateur_Esperance_limite)
colnames(resultats) <- c('Percentile',"Limite u","Espérance limitée")

library(xtable)
options(xtable.comment = FALSE)
xtable(resultats, caption = "Valeurs de l'espérance limitée pour chaque limite u donnée",
                  align = c("c", "c", "c", "c"),
                  digits = c(0, 2, 7, 7))
```

b) Intervalle de confiance pour $E[min(X,u)]$

```
# Estimer le paramètre teta de la loi exponentielle
teta <- mean(data)

# Créer 50 échantillons de 100 données
echantillion <- lapply(1:50, function(i) rexp(100, teta))

# Déterminer l'espérance limitée pour chacun des 50 échantillons et pour
#les percentiles 0.5 et 0.75
Esperance_limite_simul <- sapply(c(3,5), function(u)
                                sapply(1:50, function(j)
                                    mean(sapply(1:100, function(i)
                                        min(echantillion[[j]][i], limite_u[u])))))

# Déterminer la variance de l'espérance limitée pour les deux percentiles donnés
variance_Esperance_limite_simul <- sapply(1:2,function(i) var(Esperance_limite_simul[,i]))

# Déterminer l'intervalle de confiance
Est_Esperance_limite <- rep(0,2)
Est_Esperance_limite[1] <- Estimateur_Esperance_limite[3]
Est_Esperance_limite[2] <- Estimateur_Esperance_limite[5]

IC_Esperance_limite_simul <- lapply(1:2, function(i)
                                cbind(Est_Esperance_limite[i] - qnorm(0.975) *
                                          sqrt(variance_Esperance_limite_simul[i]),
                                    Est_Esperance_limite[i] + qnorm(0.975) *
                                        sqrt(variance_Esperance_limite_simul[i])))
```

```r
# Publier les résultats
resultats_2 <- data.frame(c(k[3], k[5]), c(limite_u[3], limite_u[5]),
                          Est_Esperance_limite, variance_Esperance_limite_simul)
colnames(resultats_2) <- c("Percentile","Limite u","Espérance limité", "Variance")

library(xtable)
options(xtable.comment = FALSE)
xtable(resultats_2, caption = "Valeurs de l'espérance limitée et de sa variance
                pour les percentiles 0.5 et 0.75",
                align = c("c", "c", "c", "c", "c"),
                digits = c(0,2,7,7,8))
```

c) $E[min(X, u)]$ **théorique**

Aucun calcul R n'a été fait dans cette section.

Question 3

a) Détermination de la fonction de survie à l'aide de l'estimateur Kaplan-Meier

```
# Présentation des données du problème
tableau1 <- {
    Temps <- c(30, 40, 57, 65, 65, 84, 90, 92, 98, 101)
    Cens <- c(1, 1, 1, 1, 0, 1, 1, 0, 1, 1)

    data.frame(Temps, Cens)
}

# Tableau détaillé permettant de calculer l'estimateur Kaplan-Meier
tableau2 <- {
    yi <- unique(tableau1[which(tableau1[,2] != 0),1])    # moments uniques des décès
    i <- 1:length(yi)
    Si <- rep(1,length(yi))                               # nombre de décès au temps yi
    ri <- c(10,9,8,7,5,4,2,1)                             # nombre de survivants au temps yi

    data.frame(i, yi, Si, ri, "Kaplan Meier" = cumprod(1-Si/ri))
}
colnames(tableau2) <- c("i", "yi", "Si", "ri", "Kaplan-Meier")

# Déterminer l'estimateur Kaplan-Meier pour chacun des moments uniques des décès
Estimateur_KM <- cumprod(1-Si/ri)
```

b) Graphique de l'estimateur Kaplan-Meier et intervalle de confiance pour $S_n(50)$

```
# Calculer un estimateur pour la variance de la fonction de survie à l'aide de
# la formule de Greenwood
formule_Greenwood <- Estimateur_KM^2 * cumsum(Si/ri/(ri-Si))

# Pour ce qui est de l'intervalle de confiance au niveau 95% pour S(50), il
# faut estimer la valeur de S(50) et de Var(S(50))

# Estimer S(50)
Sn_50 <- Estimateur_KM[2]

# Estimer Var(S(50))
Var_Sn_50 <- formule_Greenwood[2]

# Intervalle de confiance au niveau 95% pour S(50)
IC_KM <- cbind(Sn_50 - qnorm(0.975) * sqrt(Var_Sn_50),
               Sn_50 + qnorm(0.975) * sqrt(Var_Sn_50))
```

c) Intervalle de confiance pour $S_n(50)$ avec la transformation log (-log)

```
# La valeur estimée de la fonction de survie et la variance estimée sont
#toujours les mêmes que celles de la section b). Par conséquent,
# seulement l'intervalle de confiance est modifié :
```

```r
u <- exp(qnorm(0.975) * sqrt(Var_Sn_50) / Sn_50 / log(Sn_50))

IC_log_KM <- cbind(Sn_50^(1/u),
                   Sn_50^u)
```