# TP 6 Statistiques

Adib Habbou, Adel Kebli, Clark Ji

11/03/2022

## Tests d'hypothèse :

### Tests paramètriques :

#### Question 1 :

On réalise le test de Neyman-Pearson pour déterminer la région critique optimale de rejet de $H_0$ nous permettant de trancher entre nos 2 hypothèses sur la moyenne de notre loi log-normale : $\mu_0 = 0$ et $\mu_1 = 0.1$

Si $X \sim LogNormal(\mu, \sigma^2)$, X a donc pour fonction de densité :

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

Soit maintenant $X = (X_1, ..., X_n)$ avec $\forall i \in [\![1, n]\!], X_i \sim LogNormal(\mu, \sigma_0)$

Réalisons le test de Neyman-Person sur cet échantillon, on pose :

$$Z_n = \frac{L(X_1, ..., X_n, \mu_1, \sigma_0)}{L(X_1, ..., X_n, \mu_0, \sigma_0)}$$

on a donc

$$Z_n = \frac{\prod_{i=1}^{n} \frac{1}{X_i\sigma_0\sqrt{2\pi}} \exp\left(-\frac{(\ln(X_i)-\mu_1)^2}{2\sigma_0^2}\right)}{\prod_{i=1}^{n} \frac{1}{X_i\sigma_0\sqrt{2\pi}} \exp\left(-\frac{(\ln(X_i)-\mu_0)^2}{2\sigma_0^2}\right)}$$

soit donc

$$Z_n = \prod_{i=1}^{n} \exp\left(\frac{(\ln(X_i) - \mu_0)^2 - (\ln(X_i) - \mu_1)^2}{2\sigma_0^2}\right)$$

soit

$$Z_n = \prod_{i=1}^{n} \exp\left(\frac{\mu_0^2 - \mu_1^{<}2 + 2(\mu_1 - \mu_0)\ln(X_i)}{2\sigma_0^2}\right)$$

soit

$$Z_n = \exp\left(\sum_{i=1}^{n} \frac{\mu_0^2 - \mu_1^2 + 2(\mu_1 - \mu_0)\ln(X_i)}{2\sigma_0^2}\right)$$

soit

$$Z_n = \exp\left(\frac{n}{2\sigma_0^2}(\mu_0^2 - \mu_1^2) + \frac{(\mu_1 - \mu_0)}{\sigma_0^2}\sum_{i=1}^{n}\ln(X_i)\right)$$

Soit $W = \left\{Z_n > k\right\}$ la région critique optimale de rejet de $H_0$

On a alors

$$W = \left\{\frac{n}{2\sigma_0^2}(\mu_0^2 - \mu_1^2) + \frac{(\mu_1 - \mu_0)}{\sigma_0^2}\sum_{i=1}^{n}\ln(X_i) > \ln(k)\right\}$$

soit donc, car $\mu_1 > \mu_0$

$$W = \left\{\sum_{i=1}^{n}\ln(X_i) > \frac{\sigma_0^2}{(\mu_1 - \mu_0)} \times \left(\ln(k) + \frac{n}{2\sigma_0^2}(\mu_1^2 - \mu_0^2)\right)\right\}$$

soit donc

$$W = \left\{\frac{1}{n}\sum_{i=1}^{n}\ln(X_i) > \frac{\sigma_0^2}{n(\mu_1 - \mu_0)} \times \left(\ln(k) + \frac{n}{2\sigma_0^2}(\mu_1^2 - \mu_0^2)\right)\right\}$$

On pose alors comme statistique de test :

$$\boxed{T(X) = \frac{1}{n}\sum_{i=1}^{n}\ln(X_i)}$$

On sait que si $\forall i \in [\![1, n]\!], X_i \sim LogNormal(\mu, \sigma^2)$, alors $\ln(X_i) \sim \mathcal{N}(\mu, \sigma^2)$

On a donc :

$$\mathbb{E}(T(X)) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\ln(Xi)) = \mu$$

$$\mathbb{V}(T(X)) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(\ln(X_i)) = \frac{\sigma^2}{n}$$

On sait donc par théorème central limite que :

$$\frac{T(X) - \mathbb{E}(T(X))}{\sqrt{\mathbb{V}(T(X))}} = \sqrt{n}\frac{T(X) - \mu}{\sigma^2} \sim \mathcal{N}(0, 1)$$

On a alors le risque de première espèce qui vaut :

$$\mathbb{P}_{H_0}(W) = \mathbb{P}_{H_0}\left(T(X) > \frac{\sigma_0^2}{n(\mu_1 - \mu_0)} \times \left(\ln(k) + \frac{n}{2\sigma_0^2}(\mu_1^2 - \mu_0^2)\right)\right) = \mathbb{P}_{H_0}\left(T(X) > k_\alpha\right)$$

soit en utilisant le théorème central limite :

$$\mathbb{P}_{H_0}(W) = \mathbb{P}_{H_0}\left(\sqrt{n}\frac{T(X) - \mu_0}{\sigma_0} > \sqrt{n}\frac{k_\alpha - \mu_0}{\sigma_0}\right)$$

soit

$$\mathbb{P}_{H_0}(W) = 1 - \mathbb{P}_{H_0}\left(\sqrt{n}\frac{T(X) - \mu_0}{\sigma_0} < \sqrt{n}\frac{k_\alpha - \mu_0}{\sigma_0}\right)$$

Soit $\Phi$ la fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0,1)$

Or on sait que $\mathbb{P}_{H_0}(W) = \alpha$, on en déduit alors :

$$1 - \Phi\left(\sqrt{n}\frac{k_\alpha - \mu_0}{\sigma_0}\right) = \alpha$$

soit donc

$$\Phi\left(\sqrt{n}\frac{k_\alpha - \mu_0}{\sigma_0}\right) = 1 - \alpha$$

soit donc, en posant $Q_{1-\alpha}$ le quantile d'ordre $1-\alpha$

$$\sqrt{n}\frac{k_\alpha - \mu_0}{\sigma_0} = \Phi^{-1}\left(1 - \alpha\right) = Q_{1-\alpha}$$

On a donc finalement

$$\boxed{k_\alpha = \mu_0 \ + \frac{\sigma_0 \ Q_{1-\alpha}}{\sqrt{n}}}$$

De plus,

$$\beta = \mathbb{P}_{H_1}(W) = \mathbb{P}_{H_1}\left(T(X) > k_\alpha\right)$$

soit

$$\beta = \mathbb{P}_{H_1}\left(\sqrt{n}\frac{T(X) - \mu_1}{\sigma_0} > \sqrt{n}\frac{k_\alpha - \mu_1}{\sigma_0}\right)$$

soit

$$\beta = \mathbb{P}_{H_1}\left(\sqrt{n}\frac{T(X) - \mu_1}{\sigma_0} > \sqrt{n}\frac{k_\alpha - \mu_1}{\sigma_0}\right)$$

soit

$$\beta = 1 - \Phi\left(\sqrt{n}\frac{k_\alpha - \mu_1}{\sigma_0}\right)$$

soit en utilisant la valeur de $k_\alpha$ obtenue précedemment,

$$\beta = 1 - \Phi\left(\frac{\sqrt{n}}{\sigma_0} \times \left(\mu_0 \ + \frac{\sigma_0 \ Q_{1-\alpha}}{\sqrt{n}} - \mu_1\right)\right)$$

soit finalement

$$\boxed{\beta = 1 - \Phi\left(Q_{1-\alpha} + \frac{\sqrt{n}}{\sigma_0}(\mu_0 - \mu_1)\right)}$$

La valeur *alpha* représente la probabilité que notre statistique de test soit strictement supérieure à $k_{alpha}$ sachant que l'hypothèse nulle $H_0$ est vraie. Donc *alpha* correspond au risque de première espèce de notre test. On peut interpréter notre *alpha* comme la probabilité de l'erreur de première espèce, c'est-à-dire celle de décider de rejetter à tort l'hypothèse nulle $H_0$.

La valeur *beta* représente la probabilité que notre statistique de test soit strictement supérieure à $k_{alpha}$ sachant que l'hypothèse alternative $H_1$ est vraie. Donc *beta* correspond à la puissance de notre test. Par conséquent $1 - \beta$ correspond au risque de deuxième espèce de notre test. On peut interpréter notre $1 - \beta$ comme la probabilité de l'erreur de deuxième espèce, c'est-à-dire celle de décider de rejetter à tort l'hypothèse alternative $H_1$.

**Question 2 :**

```
n = 10
sigma_0 = 1
mu_0 = 0
mu_1 = 0.1
alpha = 0.1
M = 100
```

```
k_alpha <- function (n, sigma, mu, alpha) {
  q = qnorm(1 - alpha, mean = 0, sd = 1)
  return (((sigma * q) / sqrt(n)) + mu)
}
```

```
beta <- function (n, sigma, mu_0, mu_1, alpha) {
  q = qnorm(1 - alpha, mean = 0, sd = 1)
  f = pnorm(q + (sqrt(n)/sigma) * (mu_0-mu_1), mean = 0, sd = 1)
  return (1 - f)
}
```

```
T_log <- function (ech) {
  n = length(ech)
  somme = 0
  for (i in 1:n) {
    somme = somme + log(ech[i])
  }
  return(somme/n)
}
```

```
alpha_app = 0
beta_app = 0
for (i in 1:M) {
  ech0 = rlnorm(n,mu_0,sigma_0)
  ech1 = rlnorm(n,mu_1,sigma_0)
  ka = k_alpha(n,sigma_0,mu_0,alpha)

  if (T_log(ech0) > ka) {
```

```
    alpha_app = alpha_app + 1
  }
  if (T_log(ech1) > ka) {
    beta_app = beta_app + 1
  }
}

alpha_app = alpha_app / M
beta_app = beta_app / M

alpha_app
```

```
[1] 0.07
```

```
beta_app
```

```
[1] 0.11
```

Concernant le risque de première espèce on remarque bien que notre approximation de $\alpha$ tend vers la valeur 0.1 lorsqu'on augmente notre taille $M$. Le risque de choisir de rejetter $H_0$ sachant $H_0$ vraie est donc contrôlé puisqu'il ne dépasse jamais les 10.

D'autre part pour notre puissance de test $\beta$ on remarque qu'elle tend vers la valeur 0.17 lorsqu'on augmente notre taille d'échantillon $M$ cela indique que notre test est faible et que par conséquent notre test ed deuxième espèce $1-\beta$ est elevé (de l'ordre de 83) ceci implique qu'il y'a de grandes chances que l'on rejette $H_1$ sachant $H_1$ vraie.

**Question 3 :**

La valeur $p_{val}$ représente, d'après son expression, la probabilité, sachant l'hypothèse $H_0$ vraie, d'obtenir une valeur supérieure strictement à celle observée.

La valeur $p_{val}$ est donc le niveau à partir duquel on se met à rejeter l'hypothèse $H_0$. On va donc établir notre décision selon le schéma suivant :

- Si $p_{val} < \alpha$ alors on rejette l'hypothèse $H_0$ au niveau $\alpha$.
- Si $p_{val} > \alpha$ alors on conserve l'hypothèse $H_0$ au niveau $\alpha$.

Plus la valeur $p_{val}$ est faible, plus on a envie de rejeter l'hypothèse $H_0$ car cela veut dire que la valeur de la statistique utilisée pour le test est atypique pour l'hypothèse $H_0$. Cependant la valeur de $p_{val}$ n'est pas la probabilité que l'hypothèse de test soit vraie, elle indique seulement dans quelle mesure les données sont conformes à l'hypothèse de test et à ses hypothèses.

La p-value est en fait la valeur la plus petite de $\alpha$ qui permet de rejeter $H_0$, c'est la valeur critique de $\alpha$ qui fait basculer le résultat du test. Elle dépend uniquement de la réalisation de l'échantillon et permet de faire la décision avec le niveau $\alpha$ choisi arbitrairement. L'utilité de la p-value se trouve dans le fait de pouvoir prendre une décision sans être obligé de calculer la valeur critique $k_{alpha}$.

**Question 4 :**

```
n = 20
alpha_app = 0
beta_app = 0
for (i in 1:M) {
  ech0 = rlnorm(n,mu_0,sigma_0)
  ech1 = rlnorm(n,mu_1,sigma_0)
  ka = k_alpha(n,sigma_0,mu_0,alpha)

  if (T_log(ech0) > ka) {
    alpha_app = alpha_app + 1
  }
  if (T_log(ech1) > ka) {
    beta_app = beta_app + 1
  }
}

alpha_app = alpha_app / M
beta_app = beta_app / M

alpha_app
```

```
[1] 0.09
```

```
beta_app
```

```
[1] 0.17
```

```
n = 50
alpha_app = 0
beta_app = 0
for (i in 1:M) {
  ech0 = rlnorm(n,mu_0,sigma_0)
  ech1 = rlnorm(n,mu_1,sigma_0)
  ka = k_alpha(n,sigma_0,mu_0,alpha)

  if (T_log(ech0) > ka) {
    alpha_app = alpha_app + 1
  }
  if (T_log(ech1) > ka) {
    beta_app = beta_app + 1
  }
}

alpha_app = alpha_app / M
beta_app = beta_app / M

alpha_app
```

```
[1] 0.11
```

```
beta_app
```

```
[1] 0.28
```

```
n = 100
alpha_app = 0
beta_app = 0
for (i in 1:M) {
  ech0 = rlnorm(n,mu_0,sigma_0)
  ech1 = rlnorm(n,mu_1,sigma_0)
  ka = k_alpha(n,sigma_0,mu_0,alpha)

  if (T_log(ech0) > ka) {
    alpha_app = alpha_app + 1
  }
  if (T_log(ech1) > ka) {
    beta_app = beta_app + 1
  }
}

alpha_app = alpha_app / M
beta_app = beta_app / M

alpha_app
```

```
[1] 0.09
```

```
beta_app
```

```
[1] 0.38
```

On remarque que lorsque la taille de notre échantillon augmente notre valeur de $\alpha$ se rapporche encore plus de 0.1 et notre valeur de $\beta$ se rapproche également de 0.38. On peut donc dire que plus la taille de l'échantillon augmente plus la valeur approchée du risque de première espèce $\alpha$ et de la puissance de test $\beta$ se rapproche de la vraie valeur. La taille de l'échantillon a donc bel et bien une influence significative sur nos approximations, plus la taille de l'échantillon est grande plus nos approximation sont bonnes.

**Question 5 :**

Déterminons un estimateur de $\sigma^2$ pour un échantillon d'une loi log-normale

On a

$$L(X_1, ..., X_n, \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{X_i \sigma \sqrt{2\pi}} e^{-\frac{(\ln(X_i)-\mu)^2}{2\sigma^2}}$$

soit

$$L(X_1, ..., X_n, \mu, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\sum_{i=1}^{n} \frac{(\ln(X_i)-\mu)^2}{2\sigma^2}\right) \prod_{i=1}^{n} \frac{1}{X_i}$$

soit

$$\mathcal{L}(X_1, ..., X_n, \mu, \sigma^2) = -n\ln(\sigma) - \frac{n}{2}\ln(2\pi) - \sum_{i=1}^{n}\frac{(\ln(X_i)-\mu)^2}{2\sigma^2} - \sum_{i=1}^{n} X_i$$

On a alors

$$\frac{\partial \mathcal{L}}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n}\frac{(\ln(X_i)-\mu)^2}{\sigma^3} = 0$$

$$\iff \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(\ln(X_i)-\mu)^2$$

On a $\ln(X_i) \sim \mathcal{N}(\mu, \sigma^2)$

On en déduit $n\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n)$ car $\frac{\ln(X_i)-\mu}{\sigma} \sim \mathcal{N}(0,1)$

On a

$$\frac{\sqrt{n}\frac{T(X)-\mu}{\sigma}}{\sqrt{\frac{n\frac{\widehat{\sigma^2}}{\sigma^2}}{n}}} = \sqrt{n}\frac{T(X)-\mu}{\sqrt{\widehat{\sigma^2}}} \sim Student(n)$$

On peut maintenant réaliser notre test avec la même statistique de test que précedemment

$$T(X) = \frac{1}{n}\sum_{i=1}^{n}\ln(X_i)$$

On alors cette fois

$$\mathbb{P}_{H_0}(W) = \mathbb{P}_{H_0}(T(X) > k_\alpha)$$

soit

$$\mathbb{P}_{H_0}\left(\sqrt{n}\frac{T(X)-\mu_0}{\sqrt{\widehat{\sigma^2}}} > \sqrt{n}\frac{k_\alpha-\mu_0}{\sqrt{\widehat{\sigma^2}}}\right) = \alpha$$

Soit $\Psi$ la fonction de répartition de la loi de $Student(n)$ et soit $Q_\alpha^{St(n)}$ le quantile d'ordre $\alpha$ de $Student(n)$, on a alors :

$$\Psi\left(\sqrt{n}\frac{k_\alpha-\mu_0}{\sqrt{\widehat{\sigma^2}}}\right) = 1-\alpha$$

d'où

$$\sqrt{n}\frac{k_\alpha-\mu_0}{\sqrt{\widehat{\sigma^2}}} = Q_{1-\alpha}^{St(n)}$$

d'où

$$\boxed{k_\alpha = \mu_0 + \frac{\sqrt{\widehat{\sigma^2}}}{\sqrt{n}}Q_{1-\alpha}^{St(n)}}$$

De plus :

$$\beta = \mathbb{P}_{H_1}(W) = \mathbb{P}_{H_1}(T(X) > k_\alpha)$$

soit

$$\beta = \mathbb{P}_{H_1}\left(\sqrt{n}\frac{T(X) - \mu_1}{\sqrt{\widehat{\sigma^2}}} > \sqrt{n}\frac{k_\alpha - \mu_1}{\sqrt{\widehat{\sigma^2}}}\right)$$

soit

$$\beta = 1 - \Psi\left(\sqrt{n}\frac{k_\alpha - \mu_1}{\sqrt{\widehat{\sigma^2}}}\right)$$

soit

$$\beta = 1 - \Psi\left(\frac{\sqrt{n}}{\sqrt{\widehat{\sigma^2}}} \times \left(\mu_0 + \frac{\sqrt{\widehat{\sigma^2}}}{\sqrt{n}} Q_{1-\alpha}^{St(n)} - \mu_1\right)\right)$$

soit finalement :

$$\boxed{\beta = 1 - \Psi\left(Q_{1-\alpha}^{St(n)} + \frac{\sqrt{n}}{\sqrt{\widehat{\sigma^2}}}(\mu_0 - \mu_1)\right)}$$

```
estim_sigma_norm <- function (ech,  mu) {
  n = length(ech)
  somme = 0
  for (i in 1:n) {
    somme = somme + (ech[i] - mu)^2
  }
  return (sqrt( (1/(n-1)) * somme ))
}
```

```
estim_sigma_lognorm <- function (ech,  mu) {
  n = length(ech)
  somme = 0
  for (i in 1:n) {
    if (ech[i] == 0) {
      n = n - 1
    }
    else {
      somme = somme + (log(ech[i]) - mu)^2
    }
  }
  return (sqrt( (1/(n)) * somme ))
}
```

**Question 6 :**

La différence qualitative entre le comportement de l'échantillon sous l'hypothèse nulle (c'est à dire moyenne nulle) est que pour le modèle log normal notre espérance est égale à $\mathbb{E}[D_i] = e^{\sigma^2/2}$, donc cette espérance est minoré par 1 ce qui fait que l'écart entre les données obtenues en zone rurale et urbaine sera en moyenne supérieure à 1. Ce résultat semble cohérent avec ce qui est décrit par les données initiales $X_i$ et $Y_i$. On peut donc supposer que le modèle log normale décrit mieux la distrubition de nos échantillons et certaines de leurs spécificités que le modèle normale.

On réalise un test bilatéral avec les hypothèses $H_0 : \mu = \mu_0 = 0$ et $H_1 : \mu \neq \mu_0$

Pour un échantillon $D = (D_1, ..., D_n)$ où $\forall i \in [\![1, n]\!], D_i \sim \mathcal{N}(\mu, \sigma_0^2)$

On pose comme statistique de test:

$$T(D) = \frac{1}{n}\sum_{i=1}^{n} D_i$$

On a $\mathbb{E}(T(D)) = \mu$ et $\mathbb{V}(T(D)) = \frac{\sigma_0^2}{n}$

On a comme zone de rejet :

$$W = \Big\{ |T(D) - \mu_0| > K_\alpha \Big\}$$

soit

$$\mathbb{P}_{H_0}\Big( |T(D) - \mu_0| > K_\alpha \Big) = 2\Big(1 - \Phi\left(\sqrt{n}\frac{K_\alpha}{\sigma_0}\right)\Big) = \alpha$$

d'où

$$\boxed{K_\alpha = \frac{\sigma_0}{\sqrt{n}} Q_{1-\frac{\alpha}{2}}}$$

On rejette $H_0$ si

$$|T(D) - \mu_0| > \frac{\sigma_0}{\sqrt{n}} Q_{1-\frac{\alpha}{2}}$$

De plus,

$$\boxed{\beta = 1 - \Phi\Big(\frac{\sqrt{n}}{\sigma_0}(\mu_0 - \mu) + Q_{1-\frac{\alpha}{2}}\Big) + \Phi\Big(\frac{\sqrt{n}}{\sigma_0}(\mu_0 - \mu) - Q_{1-\frac{\alpha}{2}}\Big)}$$

On réalise de nouveau un test bilatéral avec les hypothèses $H_0 : \mu = \mu_0 = 0$ et $H_1 : \mu \neq \mu_0$

Pour un échantillon $D = (D_1, ..., D_n)$ où $\forall i \in [\![1, n]\!], D_i \sim \mathcal{LN}(\mu, \sigma_0^2)$

On a cette fois comme statistique de test:

$$T(D) = \frac{1}{n}\sum_{i=1}^{n} \ln(D_i)$$

La zone de rejet est alors :

$$W = \left\{ |T(D) - \mu_0| > K_\alpha \right\}$$

Par analogie avec les résultats précédents, en remplaçant $\sigma_0^2$ par son estimateur obtenu plus haut :

$$\widehat{\sigma_0^2} = \frac{1}{n} \sum_{i=1}^{n} (D_i - \mu_0)^2$$

On obtient finalement :

$$\boxed{K_\alpha = \frac{\widehat{\sigma_0^2}}{\sqrt{n}} Q_{1-\frac{\alpha}{2}}^{St(n)}}$$

et

$$\boxed{\beta = 1 - \Psi\left( \frac{\sqrt{n}}{\sqrt{\widehat{\sigma_0^2}}} (\mu_0 - \mu) + Q_{1-\frac{\alpha}{2}}^{St(n)} \right) + \Psi\left( \frac{\sqrt{n}}{\sqrt{\widehat{\sigma_0^2}}} (\mu_0 - \mu) - Q_{1-\frac{\alpha}{2}}^{St(n)} \right)}$$

```
moyenne_empirique <- function (norm) {
  somme = 0
  for (x in norm) {
    somme = somme + x
  }
  return (somme / length(norm))
}
```

```
summer_ozone <- read.csv("summer_ozone.csv")
summer_neuil <- summer_ozone$NEUIL
summer_rur <- summer_ozone$RUR.SE
summer_diff = abs(summer_neuil - summer_rur)
```

```
sigma_norm = estim_sigma_norm(summer_diff, 0)
ka_norm = (sigma_norm / sqrt(length(summer_diff))) * qnorm(0.975, 0, sigma_norm)
test_norm = moyenne_empirique(summer_diff)
ka_norm
```

```
[1] 24.22319
```

```
abs(test_norm)
```

```
[1] 12.9613
```

Pour le modèle de la loi normale, on remarque que notre statstique de test en valeur absolue est inférieure à notre $K_\alpha$. On peut en déduire qu'il ne faut pas rejetter l'hypothèse nulle $H_0$.

```
sigma_lognorm = estim_sigma_lognorm(summer_diff, 0)
ka_lognorm = (sigma_lognorm / sqrt(length(summer_diff))) * qnorm(0.975, 0, sigma_lognorm)
test_lognorm = moyenne_empirique(summer_diff)
ka_lognorm
```

```
[1] 0.5181392
```

```
abs(test_lognorm)
```

```
[1] 12.9613
```

Pour le modèle de la loi log normal, on remarque que notre statstique de test en valeur absolue est supérieure à notre $K_\alpha$. On peut en déduire qu'il faut rejetter l'hypothèse nulle $H_0$.

```
winter_ozone <- read.csv("winter_ozone.csv")
winter_neuil <- winter_ozone$NEUIL
winter_rur <- winter_ozone$RUR.SE
winter_diff = abs(winter_neuil - winter_rur)
```

```
sigma_norm = estim_sigma_norm(winter_diff, 0)
ka_norm = (sigma_norm / sqrt(length(winter_diff))) * qnorm(0.975, 0, sigma_norm)
test_norm = moyenne_empirique(winter_diff)
ka_norm
```

```
[1] 42.89973
```

```
abs(test_norm)
```

```
[1] 18.63067
```

Pour le modèle de la loi normal, on remarque que notre statstique de test en valeur absolue est inférieure à notre $K_\alpha$. On peut en déduire qu'il ne faut pas rejetter l'hypothèse nulle $H_0$.

```
sigma_lognorm = estim_sigma_lognorm(winter_diff, 0)
ka_lognorm = (sigma_lognorm / sqrt(length(winter_diff))) * qnorm(0.975, 0, sigma_lognorm)
test_lognorm = moyenne_empirique(winter_diff)
ka_lognorm
```

```
[1] 0.7241364
```

```
abs(test_lognorm)
```

```
[1] 18.63067
```

Pour le modèle de la loi log normal, on remarque que notre statstique de test en valeur absolue est supérieure à notre $K_\alpha$. On peut en déduire qu'il faut rejetter l'hypothèse nulle $H_0$.

On remarque très clairement avec le modèle normal, on est obligé de ne pas rejette l'hypothèse nulle $H_0$ ce qui nous conduit à penser qu'en moyenne la différence de nos deux échantillons est nulle. Chose qui est contradictoire avec la réalité, on sait très bien que les observations diffèrent du mileu urbain au milieu rurale.

D'autre part, avec le modèle log normal, le test de Neyman-Pearson nous conduit a rejette l'hypothèse nulle $H_0$. On est donc plus proche de la réalité de nos observations puisqu'on arrive à montrer très clairement que la moyenne de la différence de nos échantillons ne peut pas être nulle.

On conlut que le modèle log normal décrit mieux nos échantillons que le modèle normal.

**Question 7 :**

On a $D_i = X_i - Y_i$, on pose alors

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i = \frac{1}{n}\sum_{i=1}^{n} X_i - Y_i$$

On réalise un test avec les hypothèses suivantes :

$H_0$ : $D_1, ..., D_n$ symétrique autour de 0

$H_1$ : $D_1, ..., D_n$ non symétrique de 0

On a $(d_1, ..., d_n)$ en donnée

On pose $(d_1^*, ..., d_n^*) = (\pm d_1, ..., \pm d_n)$

On pose alors :

$$\overline{d^*} = \frac{1}{n}\sum_{i=1}^{n} d_i^*$$

```
summer_diff_estim = sample(c(-1, 1), size=length(summer_diff), replace=TRUE) * summer_diff
dstar = moyenne_empirique(summer_diff_estim)
cpt = 0
for (elt in summer_diff_estim) {
  cpt = cpt + 1
  if (cpt == ceiling((95/100)*length(summer_diff_estim))) {
    break
  }
}
summer_diff_sort = sort(summer_diff_estim)
summer_diff_sort[cpt]
```

```
[1] 30
```

```
abs(dstar)
```

```
[1] 1.824847
```

```
winter_diff_estim = sample(c(-1, 1), size=length(winter_diff), replace=TRUE) * winter_diff
dstar = moyenne_empirique(winter_diff_estim)
cpt = 0
for (elt in winter_diff_estim) {
  cpt = cpt + 1
  if (cpt == ceiling((95/100)*length(summer_diff_estim))) {
    break
  }
}
winter_diff_sort = sort(winter_diff_estim)
winter_diff_sort[cpt]
```

```
[1] 58
```

```
abs(dstar)
```

```
[1] 1.831533
```

On remarque que notre statistique de test $d^*$ est inférieur à notre centile $100(1-\alpha)$. Par conséquent on ne peut pas rejetter l'hypothèse nulle $H_0$, on rejette plutôt l'hypothèse alternative $H_1$. Ceci nous permet de conclure que notre échantillon est bel et bien symétrique autour de 0 à la fois pour les données en été et en hiver.

On peut donc dire que les résultats des tests non paramètriques sont cohérents avec les résultats des tests paramètriques puisque le fait que notre échantillon est symètrique autour de 0 n'est pas en contradiction avec le fait que notre échantillon suivent une loi log normale.

**Question 8 :**

On note $SN = (SN_1, ..., SN_2)$ l'échantillon des valeurs d'ozone pour la ville de Neuilly en été et $SR = (SR_1, ..., SR_2)$ celui des valeurs rurales en été.

On suppose, $\forall i \in [\![1, n]\!]$, on a $SN_i \sim \mathcal{LN}(\mu_1, \sigma)$ et $SR_i \sim \mathcal{LN}(\mu_2, \sigma)$

On réalise le test du rapport de vraisemblance avec les hypothèses suivantes :

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

On pose $\mu = (\mu_1, \mu_2)$

On pose

$$T(SN) = 2 \ln \frac{\max_{\mu \in H_1} L(SN, \mu)}{\max_{\mu \in H_0} L(SN, \mu)}$$

On a alors, d'après le quotient de vraisemblance réalisé plus haut :

$$T(SN) = 2\left( \frac{n}{2\sigma^2} (\max_{\mu \in H_0} \mu^2 - \max_{\mu \in H_1} \mu^2) + \frac{(\max_{\mu \in H_1} \mu - \max_{\mu \in H_0} \mu)}{\sigma^2} \sum_{i=1}^{n} \ln(SN_i) \right)$$

On sait que, $T(SN) \sim \chi^2(k)$ où $k = \dim(H_1) - \dim(H_0)$

On a donc ici $k = 2 - 1 = 1$

D'où $T(SN) \sim \chi^2(1)$

On rejettera alors $H_0$ si $T(SN) > Q_{1-\alpha}^{\chi^2(1)}$

```
rapport_vrais <- function(mu1,mu2,sig,ech) {
  somme = 0
  n = length(ech)
  for (i in 1:n) {
    somme = somme + (2*log(ech[i])*mu1-2*log(ech[i])*mu2+mu2^2-mu1^2)/2*sig^2
  }
  return(2*somme)
}

estim_T = optim(par = c(1, 1),
                fn = function (theta) {rapport_vrais(theta[1], theta[2], 1, summer_neuil)},
```

```
              method = "L-BFGS-B", lower = c(10^-1, 10^-1, 10^-1), control = list(fnscale = -1))$par

estim_mu1 = estim_T[1]

estim_mu2 = estim_T[2]



Q_0 = qchisq(0.975, 1)

Q_0
```

```
[1] 5.023886
```

```
T_eval = 2*log(estim_mu1/estim_mu2)

T_eval
```

```
[1] 7.568433
```

```
rapport_vrais <- function(mu1,mu2,sig,ech) {
  somme = 0
  n = length(ech)
  for (i in 1:n) {
    somme = somme + (2*log(ech[i])*mu1-2*log(ech[i])*mu2+mu2^2-mu1^2)/2*sig^2
  }
  return(2*somme)
}

estim_T = optim(par = c(1, 1),
              fn = function (theta) {rapport_vrais(theta[1], theta[2], 1, summer_rur)},
              method = "L-BFGS-B", lower = c(10^-1, 10^-1, 10^-1), control = list(fnscale = -1))$par

estim_mu1 = estim_T[1]

estim_mu2 = estim_T[2]



Q_0 = qchisq(0.975, 1)

Q_0
```

```
[1] 5.023886
```

```
T_eval = 2*log(estim_mu1/estim_mu2)

T_eval
```

```
[1] 7.60769
```

On remarque que la valeur de notre statistique de test est supérieure à notre quantile de la loi de Khi-2, on doit donc rejetter notre hypothèse nulle $H_0$ ce qui implique que $\mu_1 \neq \mu_2$. Le résultat est cohérent avec le reste, puisqu'il indique que nos observations en milieu urbain et en milieu rurale ne suivent pas la même loi log normale.

**Question 9 :**

On réalise un test de Kolmogorov-Smirnov de niveau $\alpha$ avec les hypothèses suivantes :

$H_0 : F = F_0$

$H_1 : F \neq F_0$

On estime F par rapport à la fonction de répartition empirique :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{]-\infty,x[}(X_i)$$

$F_n$ est un estimateur sans biais de F car :

$$\mathbb{E}(F_n(x)) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left(1_{]-\infty,x[}(X_i)\right) = \mathbb{P}(X_1 < x) = F(x)$$

On a alors :

$$D_n = \sqrt{n} \sup_{y \in \mathbb{R}} |F_n(y) - F(y)| \xrightarrow[n \to +\infty]{loi} W$$

où W indépendante de F qui admet comme fonction de répartition :

$$K(y) = \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2y^2)$$

On rejettera alors $H_0$ si $D_n > K^{-1}(1-\alpha)$

La statistique de test est complexe et dépend des valeurs des paramètres. Pour cela, nous pouvons faire des simulations de la statistique sous l'hypothèse nulle $H_0$ et faire une estimation par la méthode de Monte Carlo. On va simuler des observations et calculer notre quantité d'intérêt. Ceci nous fait une première réalisation de notre statistique de test. On recommence un grand nombre de fois les mêmes simulations, et on va obtenir plusieurs réalisations de la statistiqe sous l'hypothèse nulle $H_0$. On va alors estimer empiriquement notre fonction de répartition.

**Question 10 :**

On introduit les notations suivantes :

Soit $F_N$ la fonction de répartition de la loi gaussienne

En posant $F_{XY}$ la foncion de répartition de X et Y, et $F_D$ la fonction de répartition de $D = X - Y$, on pose les hypothèses suivantes pour le test de Kolmogorov concernant le premier scénario :

$H_0^{(1)} : F_{XY} = F_N$

$H_1^{(1)} : F_{XY} \neq F_N$

Ainsi que les hypothèses suivantes pour le test de Kolmogorov concernant le deuxième scénario :

$H_0^{(2)} : F_D = F_N$

$H_1^{(2)} : F_D \neq F_N$

```
options(warn=-1) # pour éviter un warning à cause de valeurs similaires
ks.test(summer_neuil, "pnorm", mean = 0, sd = 1)
```

```
    One-sample Kolmogorov-Smirnov test

data:  summer_neuil
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
ks.test(summer_rur, "pnorm", mean = 0, sd = 1)
```

```
    One-sample Kolmogorov-Smirnov test

data:  summer_rur
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
ks.test(summer_diff, "pnorm", mean = 0, sd = 1)
```

```
    One-sample Kolmogorov-Smirnov test

data:  summer_diff
D = 0.93041, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
ks.test(winter_neuil, "pnorm", mean = 0, sd = 1)
```

```
    One-sample Kolmogorov-Smirnov test

data:  winter_neuil
D = 0.97057, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
ks.test(winter_rur, "pnorm", mean = 0, sd = 1)
```

```
    One-sample Kolmogorov-Smirnov test

data:  winter_rur
D = 0.99865, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
ks.test(winter_diff, "pnorm", mean = 0, sd = 1)
```

```
    One-sample Kolmogorov-Smirnov test

data:  winter_diff
D = 0.96841, p-value < 2.2e-16
alternative hypothesis: two-sided
```

La fonction *ks.test* effectue un test de Kolmogorov-Smirnov et renvoie le nom de notre jeu de données, la valeur de notre statistique de test noté $D$, notre règle de décision avec la p-value et précise également le type de notre hypothèse alternative.

D'après les résultats de Glivenko-Kolmogorov on arrive à trouver une statistique de test qui suit asymptotiquement une distribution statistique de fonction de répartition connue, on arrive donc à utiliser le quantiel (idem que la fonction inverse de la fonction de répartition) pour établir notre région de rejet et donc notre décision.

**Question 11 :**

On introduit les notations suivantes :

Soit $F_{LN}$ la fonction de répartition de la loi log-gaussienne

En posant $F_{XY}$ la foncion de répartition de X et Y, et $F_D$ la fonction de répartition de $D = X - Y$, on pose les hypothèses suivantes pour le test de Kolmogorov concernant le premier scénario :

$H_0^{(1)} : F_{XY} = F_{LN}$

$H_1^{(1)} : F_{XY} \neq F_{LN}$

Ainsi que les hypothèses suivantes pour le test de Kolmogorov concernant le deuxième scénario :

$H_0^{(2)} : F_D = F_{LN}$

$H_1^{(2)} : F_D \neq F_{LN}$

```r
options(warn=-1) # pour éviter un warning à cause de valeurs similaires
ks.test(summer_neuil, "plnorm", mean = 0, sd = 1)
```

```
    One-sample Kolmogorov-Smirnov test

data:  summer_neuil
D = 0.99659, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```r
ks.test(summer_rur, "plnorm", mean = 0, sd = 1)
```

```
    One-sample Kolmogorov-Smirnov test

data:  summer_rur
D = 0.99747, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```r
ks.test(summer_diff, "plnorm", mean = 0, sd = 1)
```

```
    One-sample Kolmogorov-Smirnov test

data:  summer_diff
D = 0.77461, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
ks.test(winter_neuil, "plnorm", mean = 0, sd = 1)
```

```
    One-sample Kolmogorov-Smirnov test

data:  winter_neuil
D = 0.90721, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
ks.test(winter_rur, "plnorm", mean = 0, sd = 1)
```

```
    One-sample Kolmogorov-Smirnov test

data:  winter_rur
D = 0.97952, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
ks.test(winter_diff, "plnorm", mean = 0, sd = 1)
```

```
    One-sample Kolmogorov-Smirnov test

data:  winter_diff
D = 0.86633, p-value < 2.2e-16
alternative hypothesis: two-sided
```

La fonction *ks.test* effectue un test de Kolmogorov-Smirnov et renvoie le nom de notre jeu de données, la valeur de notre statistique de test noté $D$, notre règle de décision avec la p-value et précise également le type de notre hypothèse alternative.

D'après les résultats de Glivenko-Kolmogorov on arrive à trouver une statistique de test qui suit asymptotiquement une distribution statistique de fonction de répartition connue, on arrive donc à utiliser le quantiel (idem que la fonction inverse de la fonction de répartition) pour établir notre région de rejet et donc notre décision.

**Question 12 :**

Le fichier winter_ozone.csv contient 463 observations réparties sur 3 variables "date2", "NEUIL" et "RUR.SE". Il existe une différence entre les sites urbains et ruraux. La moyenne est supérieure sur les sites ruraux, de même pour la valeur maximale. D'après la figure tracée, il n'apparaît pas de différence notable entre chaque année, surtout lorsque l'on regarde la moyenne annuelle. On remarque que quasiment toutes les différences de densité sont postives, ce qui témoigne du fait que les mesures sur les sites ruraux sont bien plus élevées que celles sur les sites urbains.

En faisant l'hypothèse que l'écart-type est le même pour les quatres jeux de données, on remarque que l'on retrouve bien les bonnes moyennes empiriques. On en déduit qu'il est raisonnable de fixer le paramètre d'écart-type à la moyenne des écart-type de nos jeux de données. En effet, le calcul de l'estimateur du maximum de vraisemblance est peu sensible à la variation de ce dernier.

L'écart-type étant une mesure de la dispersion des valeurs d'un jeu de donnée, il semble logique de prendre la même valeur pour l'écart-type utilisé dans la fonction log vraisemblance de la loi normale, puisque les valeurs des écart-type des jeux de données sont assez proches les unes des autres. Cela est cohérent avec la réalité étant donnée que l'on compare des concentrations maximales en ozone dont la disparité n'est pas si différente selon la saison ou la régoin (rurale ou urbaine).

La différence est considèrable entre la log-vraisemblance des échantillons en considérant une loi normale ou une loi log normale. On trouve des valeurs bien plus grande pour la loi normale que pour la loi log normale. On préfère ainsi la loi log normale parce qu'elle colle plus aux échantillons.

En calculant les intervalles de confiance, on remarque très clairement qu'aucun intervalle de confiance ne contient zéro. Cela peut s'interpréter par le fait que nos deux paramètres $\mu$ et $\sigma$ ne valent jamais zéro. On peut ainsi dire qu'on aura jamais la même moyenne du niveau de concentration d'ozone entre une zone urbaine et une zone rurale, même si on compare pour deux mêmes saisons.

Le résultat est tout à fait cohérent avec la réalité puisque les zones urbaines sont beaucoup plus polluées que les zones rurales, l'écart en niveau de concentration d'ozone est donc logique puisque la pollution augmente la concentration d'ozone dans l'air.

D'après les tests de Kolmogorov appliquée à notre échantillon, on voit clairement que notre choix de choisir la loi log normale et beaucoup plus pertinent que celui de la loi normale. On peut donc essayer dans le futur de prédire les valeurs de concentrations d'ozone en faisant des simulations de loi log normale étant donnée les bons paramètres obtenus par estimation du maximum de vraisemblance et correspondant à des intervalles de confiance de niveau prédéfini. De plus les tests nous permettent d'affiner nos estimations, en comparant la précision de deux estimateurs afin d'avoir le modèle le plus précis possible.