

# TP Statistiques 4

Matthias LAPU | Amael Kreis

## Tests d'Hypothèse

### Tests paramétriques

1. Construire le NP test. Quelle est la statistique du test,  $T(\mathbf{X})$ ? Étant donné  $n = 10, \sigma_0 = 1, \mu_0 = 0, \mu_1 = 0.1$ , évaluer les valeurs théoriques pour  $k_\alpha$  et  $\beta$ . Quelle est l'interprétation de ces valeurs  $\alpha$  et  $\beta$ ?

Il est demandé de faire un test simple avec la méthode de Neyman-Pearson. Le test d'hypothèse est :

$$H_0 : \mu = \mu_0 \quad H_1 : \mu = \mu_1$$

D'après le cours, la région critique optimale est définie par  $W$  tel que :

$$Z_n = \frac{L(\log(X_i); \mu_1)}{L(\log(X_i); \mu_0)} = e^{\left( \frac{-\sum_{i=1}^n (\log(X_i) - \mu_1)^2 + \sum_{i=1}^n (\log(X_i) - \mu_0)^2}{2\sigma_0^2} \right)} = e^{\frac{(\mu_1 - \mu_0) \sum_{i=1}^n \log(X_i)}{\sigma_0^2}} e^{-\frac{n(\mu_1^2 - \mu_0^2)}{\sigma_0^2}}$$

$Z_n$  est une variable aléatoire continue sous  $\mathbb{P}_{\mu_0}$ . Comme dans le cours, la région critique optimale au seuil  $\alpha$  est donc :

$$W = \left\{ (\log(x_1), \dots, \log(x_n)); e^{\frac{\sum_{i=1}^n (\log(x_i) - \mu_0)^2 - \sum_{i=1}^n (\log(x_i) - \mu_1)^2}{2\sigma_0^2}} > k \right\} = \left\{ (\log(x_1), \dots, \log(x_n)); e^{\frac{(\mu_1 - \mu_0) \sum_{i=1}^n \log(x_i)}{\sigma_0^2}} e^{-\frac{n(\mu_1^2 - \mu_0^2)}{\sigma_0^2}} > k \right\}$$

Ainsi nous connaissons la statistique du test :

$$T(\mathbb{X}) = \frac{1}{n} \sum_{i=1}^n \log(x_i)$$

Afin de terminer la construction du NP test, il va falloir déterminer  $K_\alpha$ . Il va falloir se baser sur l'hypothèse  $H_0$ .

$$T(\mathbb{X}) \sim \mathcal{N}\left(\mu_0, \frac{\sigma_0^2}{n}\right)$$

En notant  $Y = \log(X)$ , nous avons donc :

$$\mathbb{P}_{H_0}\left(\frac{1}{n} \sum_{i=1}^n \log(x_i) > K_\alpha\right) = \mathbb{P}_{H_0}\left(\frac{\sqrt{n}(\bar{Y}_n - \mu_0)}{\sigma_0} > \frac{\sqrt{n}(K_\alpha - \mu_0)}{\sigma_0}\right) = 1 - \Phi\left(\frac{\sqrt{n}(K_\alpha - \mu_0)}{\sigma_0}\right) = \alpha$$

On retrouve donc  $K_\alpha$ .

$$K_\alpha = \mu_0 + \frac{\sigma_0}{\sqrt{n}} \Phi^{-1}(1 - \alpha)$$

Calculons la puissance du test.

$$\mathbb{P}_{H_1}(W) = \mathbb{P}_{H_1}(\bar{Y}_n > K_\alpha) = \beta = 1 - \Phi\left(\frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma_0}\right) + \Phi^{-1}(1 - \alpha)$$

Ainsi la construction du test est terminée.

Évaluons les valeurs théoriques.

```
#initialisation des paramètres
n<-10
sig <- 1
mu0 <- 0
mu1 <- 0.1
#calcul de k_alpha
K_alpha<-(1/sqrt(n))*(1/qlnorm(1-0.05,mu0,1/n))
print(K_alpha)
```

```
alpha<- pnorm(K_alpha,mu0,1/n)
print(alpha)
```

```
beta<-1-pnorm(K_alpha,mu0,1/n)
print(beta)
```

D'après le cours, nous savons qu' $\alpha$  est le taux d'erreur de première espèce, c'est-à-dire la probabilité d'accepter l'hypothèse 1 sachant que l'hypothèse 0 est vraie, tandis que  $\beta$  est le taux d'erreur de seconde espèce, c'est-à-dire la probabilité de garder l'hypothèse 0 sachant que l'hypothèse 1 est bonne.

2. Simulez les données avec le paramètre ci-dessus et effectuez le test de niveau  $\alpha = 0.1$   $M = 100$  fois. Donnez une approximation de  $\alpha$  et  $\beta$ . Le test contrôle-t-il les erreurs comme promis ?

```
l<-c()
for( i in 1:100){
  lnnorm<-rlnorm(10,0,1)
  l[i]<-dlnorm(lnnorm)
}
```

```
## Warning in l[i] <- dlnorm(lnnorm): le nombre d'objets à remplacer n'est pas
## multiple de la taille du remplacement
```

```
## Warning in l[i] <- dlnorm(lnnorm): le nombre d'objets à remplacer n'est pas
## multiple de la taille du remplacement
```

```
## Warning in l[i] <- dlnorm(lnnorm): le nombre d'objets à remplacer n'est pas
## multiple de la taille du remplacement
```

```
## Warning in l[i] <- dlnorm(lnnorm): le nombre d'objets à remplacer n'est pas
## multiple de la taille du remplacement
```

```
## Warning in l[i] <- dlnorm(lnnorm): le nombre d'objets à remplacer n'est pas
## multiple de la taille du remplacement
```

```
## Warning in l[i] <- dlnorm(lnnorm): le nombre d'objets à remplacer n'est pas
## multiple de la taille du remplacement
```

```
## Warning in l[i] <- dlnorm(lnnorm): le nombre d'objets à remplacer n'est pas
## multiple de la taille du remplacement
```











```
## Warning in l[i] <- dlnorm(lnnorm): le nombre d'objets à remplacer n'est pas
## multiple de la taille du remplacement
```

```
## Warning in l[i] <- dlnorm(lnnorm): le nombre d'objets à remplacer n'est pas
## multiple de la taille du remplacement
```

```
## Warning in l[i] <- dlnorm(lnnorm): le nombre d'objets à remplacer n'est pas
## multiple de la taille du remplacement
```

```
lsorted<-sort(l)
K_alpha_moy<-lsorted[90]-0.0001
beta_moy<-1-pnorm(K_alpha_moy,0,1/10)

print(K_alpha_moy)
print(alpha_moy)
```

```
## Error in print(alpha_moy): objet 'alpha_moy' introuvable
```

```
print(beta_moy)
```

On trouve des valeurs très proches même avec la marge d'erreur.

3. Au lieu de déterminer  $k_\alpha$  pour les tests, nous pouvons calculer la valeur  $p$ , définie comme

$$p - \text{val} = P_{H_0}(T(\mathbf{X}) > T(\mathbf{x}))$$

où  $T(\mathbf{x})$  est la statistique du test observée. Expliquez comment utiliser la valeur  $p$  pour établir une règle de décision pour le test.

On utilise la fonction de répartition pour trouver  $p$ :

```
lnnorm<-rlnorm(10,0,1)
p<-pnorm(dlnorm(lnnorm),0,1/10)
print(p)
```

4. Répéter pour les tailles d'échantillon croissantes  $n = 20, 50, 100$ . Quelle est l'influence de la taille de l'échantillon  $n$  sur le test?
5. Considérons le cas où  $\sigma$  est inconnu et répétez les questions précédentes. Y a-t-il une différence dans votre conclusion?

## Application: Air quality monitoring

[Airparif](#) exploite un système de surveillance de la qualité de l'air avec un réseau de sites dans la région de la capitale (Île de France) sur lesquels les mesures de la qualité de l'air sont effectuées automatiquement. Ces mesures sont utilisées pour résumer les niveaux actuels de pollution atmosphérique, pour prévoir les niveaux futurs et pour fournir des données pour la recherche scientifique, contribuant à l'évaluation des risques pour la santé et des impacts environnementaux des polluants atmosphériques.

Nous examinerons l'*ozone troposphérique* ( $O_3$ ). Ce polluant n'est pas émis directement dans l'atmosphère, mais est produit par des réactions chimiques entre le dioxyde d'azote ( $NO_2$ ), les hydrocarbures et la lumière du soleil.



Nous nous concentrerons sur les données de deux sites de surveillance: un site urbain à Neuilly-sur-seine (**NEUIL**) et un site rural (**RUR.SE**) près de la forêt de Fontainbleu.

Les données de chaque site sont des mesures quotidiennes de la concentration moyenne horaire maximale de  $O_3$  enregistrée en microgrammes par mètre cube ( $\mu g/m^3$ ), de 2014 à 2019 inclusivement. Pour nous concentrer sur la question de la saison, nous comparons les données de *hiver* (novembre-février inclus) (`Ozone_hiver.csv`) et *été* (mai - août inclus) (`Ozone_ete.csv`).

Nous souhaitons savoir comment la distribution des mesures de l’ozone varie-t-elle entre les sites urbains et ruraux. Nous désignons les données sur l’ozone du site urbain par  $X_i$  et le site rural par  $Y_i$ ,  $i = 1, \dots, n$ , l’indice indiquant les  $n$  jours différents pour lesquels nous avons des mesures et définissons la variable  $D_i = X_i - Y_i$  pour la différence.

6. Appliquer l’analyse exploratoire des données (TPs 1-2) et suggérer un modèle approprié pour  $D_i$ .
7. En supposant que les différences  $D_i$  forment un échantillon iid suivant une loi normale  $N(\mu, \sigma^2)$ , quelle est l’hypothèse sous-jacente que nous voulons tester ? Définir  $H_0$  et  $H_1$  et effectuez le test pour les données en été et en hiver séparément. Quelle est la conclusion?

## Méthodes de simulation pour les tests d’hypothèse

Un test d’hypothèse valide exige que nous rejetions incorrectement l’hypothèse nulle une proportion appropriée du temps (par exemple, au plus 5% de fois).

Ainsi pour une statistique de test donnée,  $T(\mathbf{X})$  telle que nous rejeterons l’hypothèse nulle si  $T(\mathbf{X})$  est plus grande (ou plus petite) qu’un certain seuil, si nous voulons un test de taille  $\alpha$  nous devons pouvoir calculer  $k_\alpha$  tel que

$$\Pr(T(\mathbf{X}) > k_\alpha | \text{Hypothèse nulle vraie}) = \alpha.$$

Si nous ne pouvons pas calculer  $k_\alpha$  analytiquement, nous pouvons utiliser la simulation pour aider à choisir un  $k_\alpha$  approprié. Ce que nous devons faire est de simuler des ensembles de données répliqués sous l’hypothèse Nulle.

### Monte Carlo approximation to the sampling distribution of the test statistic

- (1) Simuler des  $M$  ensembles de données indépendants,  $\mathbf{x}_1, \dots, \mathbf{x}_M$  sous l’hypothèse nulle.
- (2) Pour chaque  $\mathbf{x}_i$ , calculez la statistique de test  $T(\mathbf{x}_i)$ .
- (3) Les valeurs  $\{T(\mathbf{x}_1), \dots, T(\mathbf{x}_M)\}$  sont les échantillon de la loi de  $T(\mathbf{X})$ , desquelles on obtient la fonction de répartition empirique comme l’approximation de Monte Carlo.
- (4) Estimez le seuil  $k_\alpha$  comme le centile empirique  $100(1 - \alpha)$  de cette approximation.

Notez que pour implémenter cela, nous avons seulement besoin de pouvoir simuler des données sous l’hypothèse nulle. Dans de nombreuses applications scientifiques, les scientifiques effectueront un test d’hypothèse en utilisant cette approche. Ils/Elles décideront d’une statistique du test, en fonction de leur compréhension du problème, et utiliseront la simulation pour calculer la valeur seuil.

### Test d’ajustement de Kolmogorov

Nous pouvons appliquer la stratégie de simulation pour évaluer la pertinence des modèles statistiques pour les données.

Soit  $X_1, \dots, X_n$  un échantillon de loi inconnue  $P_\theta$  de fonction de répartition  $F$  supposée continue. L’objectif du test de Kolmogorov est l’ajustement de la loi inconnue  $P$  à une loi connue  $P_0$  de fonction de répartition continue  $F_0$ :

$$H_0 : F = F_0 \quad H_1 : F \neq F_0$$

8. Construire le test de Kolmogorov (Kolmogorov-Smirnov) de niveau  $\alpha$  (sur la base de l'approximation asymptotique). Suggérer une méthode alternative basée sur la simulation.

Supposons que le modèle le mieux ajusté ait la valeur de paramètre  $\hat{\theta}$ . Soit  $F_0$  la fonction de répartition du modèle ajusté  $P_{\hat{\theta}}$ .

9. Pour les données sur l'ozone, nous voulons tester si l'hypothèse de gaussianité était appropriée. Nous envisageons deux scénarios. Le premier est que les données originales de l'ozone suivent une loi gaussienne ( $H_0^{(1)}$ ). Le second suppose que seules les différences suivent la loi gaussienne ( $H_0^{(2)}$ ). Effectuez les tests utilisant la méthode asymptotique et la méthode de simulation. Résumez vos conclusions.