TP5 Statistiques noté : Test d'hypothèse statistique

Nom1Prenom1-Nom2Prenom2

9 mai 2025

A. Tests d'hypothèses

1. Tests paramétriques

Pour les échantillons iid $X_i \sim \mathcal{N}(\mu, \sigma^2), i = 1, \ldots, n$, *considerons un test d'hypothèse simple*

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

où $\mu_1 > \mu_0$ *et* $\sigma = \sigma_0$ *est connu. Pour* $\mathbf{X} = (X_1, \ldots, X_n)$ *donné, nous savons que le test de Neyman-Pearson (NP) rejeté* H_0 *si*

$$T(\boldsymbol{X}) > k_\alpha$$

pour une valeur seuil appropriée k_α. *On rappelle que*

$$\alpha = P_{H_0}(T(\mathbf{X}) > k_\alpha) \quad \beta = P_{H_1}(T(\mathbf{X}) > k_\alpha)$$

Celles-ci donnent des garanties théoriques pour contrôler les erreurs de décision, α *et* $1 - \beta$.

Ex1. Nous voulons construire un test NP. Donner la statistique de test $T(\mathbf{X})$.

Ex2. Étant donné $n = 50, \sigma_0 = 1, \mu_0 = 0, \mu_1 = 0.1$, évaluer les valeurs théoriques pour k_α et β. Quelle est l'interprétation de ces valeurs α et β?

Ex3. Simuler des données avec le paramètre ci-dessus et effectuer le test de niveau $\alpha = 0.1$ $M = 100$ fois. Donnez un approxmation de α et β. Le test contrôle-t-il l'erreurs comme promis ?

```
alp <- 0.1
nsimu <- 100
n <-50
mu0 <- 0
mu1<-0.1
sig0 <- 1
Kalpha <- ...
beta <- ...
N_alpha <- c()
N_beta <- c()
for(simu in 1:nsimu){
  X0 <- rnorm(n,mu0,sig0)
  X1 <- rnorm(n,mu1,sig0)
  N_alpha <- append(N_alpha,mean(X0)>Kalpha)
  N_beta <- append(N_beta,mean(X1)>Kalpha)
}
alpha_estim <- mean(N_alpha)
beta_estim <- mean(N_beta)
print(c("estmation de alpha :",alpha_estim))
print(c("estmation de beta :",beta_estim))
```

Comment peut-on interpreter le plot suivant ?

```
alphas <- seq(0,1,0.1)
betas <- pnorm((mu1-mu0)*sqrt(n)/sig0 + qnorm(alpha,0,1),0,1)
plot(alphas,betas,ylab = "beta",xlab="alpha")
```

Ex4. Au lieu de déterminer k_α pour les tests, nous pouvons calculer la valeur p, définie comme

$$p_{val} = P_{H_0}(T(\mathbf{X}) > T(\mathbf{x}))$$

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

Ex5. Consideron le cas où σ est inconnu, quell est la statistique de test ? Y a-t-il une différence dans votre conclusion?

2. Construction d'un test statistique par simulation

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).

Losque nous disposons d'une statistique de test $T(\mathbf{X})$, et que nous souhaitons rejecter l'hypothèse nulle H_0 si $T(\mathbf{X})$ est plus grande (ou plus petite) qu'un certain seuil, alors pour avoir un test de niveau α, il faut calculer le seuil k_α tel que

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

Cependant, si nous ne pouvons pas calculer k_α de manière analytique (c'est-à-dire via une formule exacte), nous pouvons utiliser la simulation pour estimer k_α. Ce que nous devons faire est de simuler des ensembles de données répliqués sous l'hypothèse Nulle.}

Noter que pour implémenter un test d'hypothèse basé sur la simulation, il suffit de pouvoir simuler des données sous l'hypothèse nulle. On n'a pas besoin de connaître analytiquement la loi de la statistique de test. Dans de nombreuses disciplines scientifiques, cette approche est très utilisée. Les chercheurs choisissent une statistique de test qui réflète bien l'effet ou le phénomène qu'ils cherchent à détecter et utilisent la simulation pour calculer la valeur seuil à partir de la distribution empirique de la statistique choisie sous l'hypothèse nulle.

Test d'ajustement de Kolmogorov

Nous pouvons appliquer la stratégie de simulation pour évaluer la pertinence des modèles statistiques vis-à-vis des données observées.

Soit $X_1, \ldots, X_n$ un échantillon de loi inconnue $P_{\boldsymbol{\theta}}$ 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:

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

Ex6. Supposons que $P_0 = \mathcal{N}(\mu, \sigma)$ avec $\theta = (\mu, \sigma)$ sont connus, construire le test de Kolmogorov (Kolmogorov-Smirnov) de niveau α (sur la base de l'approximation asymmtotique).

```
### Remplir la fonction suivante pour calculer la statistique de test Kolmogorov
ks_stat <- function(x, mu, sigma) {
  x_ord <- sort(x)
  F_emp <- (1:n) / n # calculer la distribution empirique
  F_theo <- pnorm(..., mean = mu, sd = sigma) # calculer la distribution theorique
  D <- ... #calcule de la KS distance
  return(...)
}
```

```r
## Prenez mu et sigma a votre choix
mu = ...
sigma = ...
x_obs = ...  # générer une observation de modèle choisi.
D_obs <- ks_stat(x=..., mu= ..., sigma=...) # calculer la KS statistique
###les valeurs critiques
alp = 0.05
c_alpha = 1.36 ## le quantile (1-alpha) de la loi Kolmogorov
k_a = ...
###Conclusion
if (D_n > k_a) {
  cat("On rejette H0 : D_n =", D_n, "> valeur critique =", k_a, "\n")
} else {
  cat("On ne rejette pas H0 : D_n =", D_n, "<= valeur critique =", k_a, "\n")
}
```

Ex7. Utiliser la fonction `ks.test()` en `R` pour éffectuer le test, puis comparer les résultats avec ceux obtenus dans l'exercice précédent.

Ex8. Supposons que le modèle le mieux adjusté ait la valeur de paramètre $\hat{\boldsymbol{\theta}}$. Soit F_0 la fonction de répartition du modèle ajusté $P_{\hat{\boldsymbol{\theta}}}$.

Notons que lorsque $\hat{\theta} = (\hat{\mu}, \hat{\sigma})$ est estimé à partir des données, l'approximation asymptotique standard de Kolmogorov–Smirnov n'est plus valide. La distribution limite de la statistique de Kolmogorov–Smirnov dépend de la méthode d'estimation et du modèle sous-jacent. Nous voulons donc proposer une méthode alternative.

Construire un test Kolmogorov par une méthode alternative basée sur la simulation.

```r
mu_hat <-...
sigma_hat <- ...
simu <- 1000
D_sim <- numeric(simu)

for (s in 1:simu) {
  # Générer des données sous H0
  x_sim <- rnorm(n, mean = mu_hat, sd = sigma_hat)

  # Réestimer les paramètres sur l'échantillon simulé
  mu_sim <- ...
  sigma_sim <- ...

  # Calculer la stat KS
  D_sim[s] <- ks_stat(x=..., mu=..., sigma=...)
}

## Calcul de la valeur critique et décision
quantile_alpha <- quantile(x=..., probs= ...)

## Conclusion
....
```

B. 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 (Ile 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, $i = 1, \ldots, n$, 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 difference.

Télécharger les fichiers `Ozone_hiver.csv` et `Ozone_ete.csv` et les placer dans la même répétoire de votre espace de travail courant. Utiliser les codes suivant pour impotorter les données

```
setwd("chemin/vers/le/dossiers")
Hiver = read.csv("Ozone_hiver.csv")
Ete = read.csv("Ozone_ete.csv")

head(Hiver)
head(Ete)
```

```
X_hiver = Hiver[,'NEUIL']; Y_hiver = Hiver[,'RUR.SE']
X_ete= Ete[,'NEUIL']; Y_ete = Ete[,'RUR.SE']
```

Ex9. Appliquer l'analyse exploratoire des données (TPs 1-2) et suggérer un modèle approprié pour D_i.

Ex10. 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 lest pour les données en été et en hiver séparément. Quelle est la conclusion?

Ex11. (Bonus) 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 ($H_0^{(2)}$). Effectuer les tests utilisant la méthode asymptotique et la méthode de simulation. Résumez vos conclusions.