STA211

Médous Charles, Koen Aristote

4/9/2020

Partie I

Q1

Sous H_0, le vecteur $(X_1, ..., X_n)$ est i.i.d à valeurs réelles.

- Montrons tout d'abord que les (X_i) sont $\mathbb{P}$-p.s distincts:
  soit $i \neq j$,

$$\begin{aligned}
\mathbb{P}(X_i = X_j) &= \int_{\mathbb{R}^2} \delta_{x=y} f(x) f(y) \mu(\mathrm{d}x\mathrm{d}y) \\
&= \int_{\mathbb{R}} f(x) \left(\int_{\mathbb{R}} \delta_{y=x} f(y) \mu(\mathrm{d}y) \right) \mu(\mathrm{d}x) \qquad \text{(Fubini)} \\
&= \int_{\mathbb{R}} f(x) \mu(\{f(x)\}) \mu(\mathrm{d}x) \\
&= 0
\end{aligned}$$

  Maintenant, on regarde:

$$0 \leq \mathbb{P}(\exists (i \neq j) / X_i = X_j) = \mathbb{P}(\cup_{i \neq j} (X_i = X_j)) \leq \sum_{i \neq j} \mathbb{P}(X_i = X_j) \leq 0$$

- Donc on sait que les R_i prennent toutes les valeurs entières sur $\{1, ..., n\}$ et l'on va pouvoir confondre le vecteur aléatoire $R = (R_1, ..., R_n)$ avec la permutation associée. R est donc à valeur dans $\sum_n$ (ensemble des permutations de $\{1, ..., n\}$).
  Sous l'hypothèse H_0, les X_i sont indépendants et de même loi donc ils sont interchangeables, donc toutes les permutations sont possibles et équiprobables.
  Par conséquent R suit une loi uniforme sur $\sum_n$ avec la proba $Card(\sum_n) = \frac{1}{n!}$
- On va maintenant montrer que les marginales R_k suivent une loi uniforme.
  Posons

$$\pi_k \quad : \left| \begin{array}{ccc} \sum_n & \to & \{1, ..., n\} \\ \sigma & \mapsto & \sigma(k) \end{array} \right.$$

  Tout les éléments de $\{1, ..., n\}$ ont $(n-1)!$ antécédents par l'application π_k.

$$\begin{aligned}
\mathbb{P}(R_k = p) &= \mathbb{P}(\pi_k(R) = p) \\
&= \mathbb{P}(R \in \pi_k^{-1}(p)) \\
&= \mathbb{P}(\bigcup_{\sigma/\pi_k(\sigma)=p} (R = \sigma)) \qquad \text{réunion disjointe de } (n-1)! \text{ éléments équiprobables} \\
&= \mathbb{P}(R = \sigma_0)(n-1)! \\
&= \frac{(n-1)!}{n!} \\
&= \frac{1}{n} \qquad \text{indépendant de p}
\end{aligned}$$

Donc la loi de R_k est bien une uniforme sur $\{1, ..., n\}$

- Soit $k \neq l$; on cherche la loi de R_k sachant R_l.
  A R_l fixé (un entier de $\{1, ..., n\}$), les $(R_i)_{i \neq l}$ prennent toutes les valeurs de $\{1, ..., n\} \backslash R_l$
  $\widetilde{R}^l = (R_1, ..., R_l, ..., R_n)$ est toujours une permutation de $\sum_n$ mais à R_l fixé. $\widetilde{R}^l$ est en fait une application sur l'ensemble $\{\sigma \in \sum_n \quad | \quad \pi_l(\sigma) = R_l\}$, de cardinal $(n-1)!$
  Puis, comme précédemment, on montre que:

$$\forall k \neq l, p \neq m, \qquad \mathbb{P}(R_k = p \quad | \quad R_l = m) = \frac{1}{n-1}$$

Q2

La variable aléatoire S(i,j) représente la somme des rangs des variables aléatoires $X_i, \ldots, X_j$, c'est à dire entre les temps i et j. S(i,j) prendra de grandes valeurs dans le cas où la valeur des variables aléatoires $X_i, \ldots, X_j$ auront de plus fortes valeurs relativement aux variables aléatoires $X_k, k \in \{1, \ldots, i-1, j+1, \ldots, n\}$. Sou sl'hypothèse H_0 on va s'attendre à ce que cette quantité ne dépende que de $j - i$, i.e la largeur de la fenêtre et pas la position de cette fenêtre.

Q3

$$m_{i,j} := \mathbb{E}\left[S(i,j)\right] = \sum_{k=i}^{j} \mathbb{E}\left[R_k\right] = \sum_{k=i}^{j} \frac{n+1}{2} = \frac{1}{2}\left(j - i + 1\right)\left(n + 1\right)$$

Car R_k suit une loi uniforme discrète sur $\{1, \ldots, n\}$

Q4

- Calculons tout d'abord, pour $k \neq l$, $\text{cov}(R_k, R_l)$:

$$\begin{aligned}
cov(R_k, R_l) &= \sum_{i=1}^{n}\sum_{j=1}^{n} ij\ \mathbb{P}(R_k = i \cap R_l = j)\ -\ \mathbb{E}(R_k)\mathbb{E}(R_l) \\
&= \sum_{i\neq j} ij\ \mathbb{P}(R_k = i | R_l = j)\mathbb{P}(R_l = j)\ -\ \mathbb{E}(R_k)^2 \\
&= \sum_{i\neq j} ij\ \frac{1}{n-1}\frac{1}{n}\ -\ \left(\frac{n+1}{2}\right)^2 \\
&= \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} ij\ -\ \sum_{i=1}^{n} i^2\right]\ -\ \frac{(n+1)^2}{4} \\
&= \frac{1}{n(n-1)}\left[\left(\frac{n(n+1)}{2}\right)^2\ -\ \frac{n(n+1)(2n+1)}{6}\right]\ -\ \frac{(n+1)^2}{4} \\
&= \frac{n+1}{n-1}\left\{\frac{n(n+1)}{4}\ -\ \frac{2n+1}{6}\ -\ \frac{(n+1)(n-1)}{4}\right\} \\
&= \frac{n+1}{12(n-1)}\left\{3n(n+1)\ -\ 2(2n+1)\ -\ 3(n^2-1)\right\} \\
&= -\frac{n+1}{12} \qquad \text{indépendant de } k \text{ et de } l
\end{aligned}$$

- Calculons maintenant, $\text{var}(R_k)$:

$$\begin{aligned}
var(R_k) &= \mathbb{E}(R_k^2)\ -\ \mathbb{E}(R_k)^2 \\
&= \sum_{i=1}^{n} \frac{i^2}{n}\ -\ \left(\frac{n+1}{2}\right)^2 \\
&= \frac{n(n+1)(2n+1)}{6n}\ -\ \frac{(n+1)^2}{4} \\
&= \frac{n^2-1}{12} \qquad \text{indépendant de } k
\end{aligned}$$

- On peut maintenant facilement calculer $\mathbb{V}(S(i,j))$:

$$\begin{aligned}
\mathbb{V}\left(S(i,j)\right) &= \mathbb{V}(\sum_{k=i}^{j} R_k) \\
&= \sum_{k=i}^{j} var(R_k) \; + \sum_{i \le k \ne l \le j} cov(R_k, R_l) \\
&= (j-i+1)\frac{n^2-1}{12} \; - \; \frac{n+1}{12} \sum_{i \le k \ne l \le j} \\
&= (j-i+1)\frac{n^2-1}{12} \; - \; \frac{n+1}{12}(\sum_{i,k} - \sum_{i}) \\
&= (j-i+1)\frac{n^2-1}{12} \; - \; \frac{n+1}{12}((j-i+1)^2 - (j-i+1)) \\
&= \frac{(j-i+1)(n+1)}{12}\left\{n-1 \; - \; (j-i+1-1)\right\} \\
&= \frac{(j-i+1)(n+1)(n-j+i-1)}{12}
\end{aligned}$$

Partie 2

Q5

On crée une fonction simulant une réalisation de T_n:

simul_T(n):

On simule un échantillon R de taille n obtenu par tirage sans remise de n valeurs dans $\{1, ..., n\}$.

Ensuite on crée la matrice S: $S_{i,j} = (1_{i \le j})(R'_{i \le j})$ le produit matriciel des parties triangulaires superieures de $R' \in \mathcal{M}_n(\mathbb{R})$ dont chaque ligne est égale à l'échantillon R et de la matrice 1 une matrice de taille nxn composée de 1.

mij = matrix(nxn)

for i in 1:n for j in i:n

```
mij[i,j] = 0.5*(n+1)*(j-i+1)
shape(S)
```

end end vij= ones(nxn)

for i in 1:n for j in i:n

```
vij[i,j] = (1/12)*(j-i+1)*(n-j+i-1)
```

end end Tij = upper.tri((S - mij)/sqrt(vij)) return(Tij) end

Ainsi on obtient une réalisation de T_n. On peut ensuite répéter cette simulation M fois et tracer l'histogramme afin de représenter la densité sous H0.

```
rm(list = objects())
n=10

M_ij = function(n)
{
    Mij = matrix(0,n,n)
```

```
  for(i in c(1:n))
  {
    for(j in c(i:n))
    {
      Mij[i,j] = 0.5*(n+1)*(j-i+1)
      }
  }
    return(Mij)
}

V_ij = function(n)
{
  Vij = matrix(1,n,n)
  for(i in c(1:n))
  {
    for(j in c(i:n))
    {
      Vij[i,j] = (1/12)*(j-i+1)*(n-j+i-1)*(n+1)
      }
  }
  return(Vij)
}

simul_Tn = function(n,Mij_,Vij_,getindex = 0)
{
  R = sample.int(n,n,replace=FALSE)

  Rmat = rbind(R)[rep(1,n), ]
  Rmat[lower.tri(Rmat)]=0
  onetriup = matrix(1,n,n)
  onetriup[lower.tri(onetriup)]=0
  Sij = Rmat%*%onetriup

    Tij = (Sij - Mij)/sqrt(Vij)
    Tij[1,] = 0
    Tij[,n] = 0
    Tmax = max(Tij)
    indices = which(Tij == max(Tij),arr.ind = T)
    indices=as.vector(indices)
    if(getindex == 0){
    return(Tmax)} else {return(list(Tmax,indices))}
}

get_h0_sample = function(n,M,Mij__,Vij__)
{
  echants_vec = rep(n,M)
  echants = lapply(echants_vec,simul_Tn,Mij_ = Mij__,Vij_=Vij__,getindex=0)
  return(unlist(echants))
}
```

Q6

On génère donc un large échantillon iid de taille M de T_n en faisant appel M fois à la fonction simulTn(n) afin de représenter la densité. On pourrait extraire le quantile empirique d'ordre $1-\alpha$ de l'échantillon puis comparer avec la valeur de la statistique $T_n(x)$ sachant que pour de larges valeurs de M on s'approcherait de la valeur réelle du quantile. Mais on ne pourrait pas appliquer de méthode de Monte Carlo pour l'estimer et on ne pourrait pas facilement déduire un intervalle de confiance de la valeur de ce quantile. Au vu de la question posée ensuite nous procéderons comme suit

On calcule la valeur de la statistique $T_n(x) = t^*$ à partir de l'écantillon $x_{1:n}$. Pour cela :

- On calcule les matrices m_{ij} et v_{ij}
- On prend le classement des valeurs de l'échantillon. On obtient donc une permutation
- On assigne à une variable R le vecteur contenant les rangs des valeurs de l'échantillon.
- On calcule ensuite la statistique en calculant la matrice S puis T à partir de S,m,v.
- On prend finalement la valeur maximale de cette matrice triangulaire superieure ce qui nous donne t^(étoile) à partir de R comme dans la fonction simulTn.

On procède à estimer la probabilité $\mathbb{P}(T_n > t^*)$ c'est à dire estimer l'espérance de la variable aléatoire $Y = 1_{T_n > t^*}$ que l'on notera Z et que l'on peut estimer par $\overline{Y} = \frac{1}{M}\sum_{i=1}^{M} Y_i$. on peut ensuite calculer un intervalle de confiance sur cette estimation comme on le verra par la suite. Si les bornes de cet intervalle sont inferieures à α on peut alors conclure que $T_n > t_{1-\alpha}$ car $\mathbb{P}(T_n > t_{1-\alpha}) = 1 - \mathbb{P}(T_n \leq t_{1-\alpha}) = \alpha$ et que la fonction de répartition est croissante donc $(\mathbb{P}(T_n > t_{1-\alpha}) > \mathbb{P}(T_n > t^*)) \implies T_n = t^* > t_{1-\alpha}$

Ainsi on peut conclure que l'on rejette H0 dans le cas décrit ci-dessus. Dans ce cas on a donc $T_n(x) > t_{1-\alpha}$ on peut donc considérer que la periode (i_{max}, j_{max}) présente des températures anormalement élevées.

Partie 3

Q7, Q8

```
get_tn= function(x_ech,Mij_,Vij_,getindex=0)
{
  n_ = length(x_ech)
  R = rank(x_ech) #rang des valeurs de x_ech
    Rmat = rbind(R)[rep(1,n_), ] #matrice ac chaque ligne =R
  Rmat[lower.tri(Rmat)]=0  #partie triangulaire inf =0
  onetriup = matrix(1,n_,n_)
  onetriup[lower.tri(onetriup)]=0 #matrice triangulaire sup = 1
  Sij = Rmat%*%onetriup #On calcule Sij

    Tij = (Sij - Mij)/sqrt(Vij) #on calcule Tij
    Tij[1,n] = 0 #T1n=0
    Tmax = max(Tij) #on prend le max
    indices = which(Tij == max(Tij),arr.ind = T)
    indices=as.vector(indices) #on prend l'argmax
    if(getindex == 0){
    return(Tmax)} else {return(list(Tmax,indices))}

}

get_h0_sample = function(n,M,Mij__,Vij__) #fonction générant l'échantillon
```

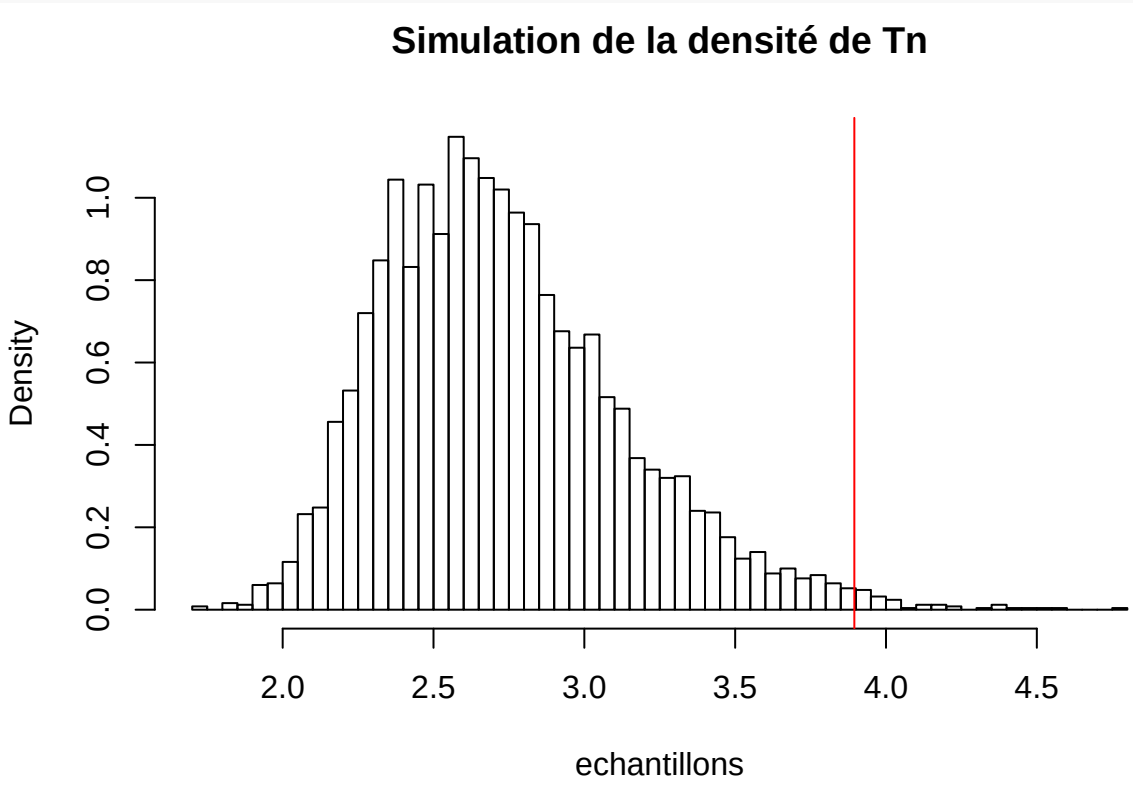
```
{
  echants_vec = rep(n,M)
  echants = lapply(echants_vec,simul_Tn,Mij_ = Mij__,Vij_=Vij__,getindex=0)
  return(unlist(echants))
}


#import des données Hobart
hobart <- read.table("https://papayoun.github.io/courses/2020_monte_carlo/enonces_td/hobart.txt",
                     sep = ";", header = TRUE)

Temp= hobart$mean_january_temperature
n = length(Temp)
Mij = M_ij(n)
Vij=V_ij(n)

t_star= get_tn(Temp,Mij,Vij,getindex = 1) #Calcul de Tn à partir de gettn
t_star_val = t_star[[1]] #Valeur de la statistique Tn
t_star_index= t_star[[2]] #argmax Tij
M = 5000
echantillons = get_h0_sample(n,M,Mij,Vij) #on génère un M échantillon iid de T_n

hist(echantillons,proba=T,breaks = 100,main = "Simulation de la densité de Tn") # on représente l'histo
abline(v=t_star_val,col='red') #on affiche où se situe la valeur de la statistique
```




```
hobart$year[t_star_index] #on affiche la fenêtre correspondant aux indices obtenus.
```

```
## [1] 2003 2017
```

```
t_star_val #on affiche la valeur de la statistique
```

```
## [1] 3.895098
```

```
##[1] 2003 2017
##[1] 3.895098
```

Q9,10

La fenêtre sur laquelle on calcule la statistique est donc 2003-2017. Soit $Y_n = 1_{T_n > t^*}$ On a:

$$\mathbb{P}(T_n > t^*) = \mathbb{E}[1_{T_n > t^*}] = Z$$

Les réalisations des Y_i correspondant à des variables aléatoires iid suivant une loi de Bernoulli de paramètre Z, d'espérance finie, nous pouvons simuler cette probabilité par méthode de Monte-Carlo car par la loi des grands nombres:

$$\overline{Y} = \frac{1}{M}\sum_{i=1}^{M} 1_{T_i > t^*} \xrightarrow{P} Z$$

On obtient donc l'estimation de la probabilité Z:

```
mean(as.numeric(echantillons>t_star_val)) ##estimation de la probabilité
```

```
## [1] 0.0088
```

Les réalisations de notre échantillon Y étant issues de variables aléatoires iid suivant une loi de Bernoulli, et donc de moments d'ordre 1 et 2 finis, par le TCL:

$$\frac{\sqrt{M}(\overline{Y} - Z)}{\sqrt{\mathbb{V}[Y]}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Ne connaissant pas la variance de Y nous pouvons l'estimer par un estimateur consistant de la variance, la variance empirique biaisée:

$$\hat{\sigma}_Y^2 = \frac{1}{M}\sum_{i=1}^{M} \left(Y_i - \overline{Y}\right)^2$$

Ainsi en remplacant la variance par son estimateur consistant, par le lemme de Slutzky on obtient un IC de niveau alpha qui asymptotiquement correspond à l'intervalle de confiance de la moyenne avec variance connue :

$$\begin{aligned}\mathbb{P}\left(\left|\frac{\sqrt{M}(\overline{Y} - Z)}{\sqrt{\hat{\sigma}_Y^2}}\right| \leq q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)}\right) &= \mathbb{P}\left(-q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \leq \frac{\sqrt{M}(\overline{Y} - Z)}{\sqrt{\hat{\sigma}_Y^2}} \leq q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)}\right) \\ &= \mathbb{P}\left(\overline{Y} - \frac{q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)}\sqrt{\hat{\sigma}_Y^2}}{\sqrt{M}} \leq Z \leq \overline{Y} + \frac{q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)}\sqrt{\hat{\sigma}_Y^2}}{\sqrt{M}}\right) = 1 - \alpha\end{aligned}$$

Ainsi on calcule l'IC de niveau 5%:

```
q975 = qnorm(0.975)
Y_barre = mean(echantillons>t_star_val)
sigmahat = sqrt(var(echantillons>t_star_val)*(M-1)/M)
IC = c(Y_barre - (q975*sigmahat/sqrt(M)), Y_barre + (q975*sigmahat/sqrt(M)))
IC
```

```
## [1] 0.00621128 0.01138872
```

On peut donc conclure que 95% des fois sera dans cet IC et donc que $0.005 \leq P(T_n > t^*) \leq 0.010$

De plus les bornes de cet interval de confiance étant inferieures à 0.05 on en conclut que $T_n > q_{0.95}^{T_n}$ et donc on rejette l'hypothèse nulle ce qui nous permet d'en conclure que la fenêtre sur laquelle la statistique T_n est calculée contient des températures qui sont anormalement élevées en Tasmanie. Ainsi les températures moyennes observées à Hobart entre 2003 et 2017 sont anormalement élevées.