



Sciences
Technologies
Santé

2D

Travaux dirigés

L'attribution d'événements extrêmes

Corrigé

Auteur :



UNIVERSITÉ
PARIS-EST CRÉTEIL
VAL DE MARNE



Réalisation :



Prérequis : bases de la statistique inférentielle, méthode du maximum de vraisemblance, information de Fisher, calcul différentiel. Des connaissances sur la convergence en loi de l'estimateur du maximum de vraisemblance, la delta-méthode et les vecteurs gaussiens seront utiles, même si l'énoncé procure le nécessaire à ce sujet. Une connaissance minimale du logiciel R serait un plus.

Résumé

Dans cette activité, on découvrira le principe et la mise en œuvre d'une méthode possible d'Attribution d'Événements Extrêmes (EEA). Cette approche consiste à comparer le risque d'un événement extrême entre deux mondes, l'un dans lequel les activités humaines sont présentes et l'autre non. La première partie de l'activité introduit le modèle mathématique et est colorée statistique mathématique, tandis que la seconde partie est dédiée à la mise en œuvre de la méthode à l'aide du logiciel R et de données de températures issues de modèles climatiques.

Introduction : le contexte scientifique

Pour bien débuter, voici des extraits du Chapitre 11 du 6ème rapport du GIEC¹ qui introduisent bien la problématique de l'attribution des événements extrêmes :

Il est établi que les émissions de gaz à effet de serre d'origine humaine ont entraîné une augmentation de la fréquence et/ou de l'intensité de certains phénomènes météorologiques et climatiques extrêmes depuis l'époque préindustrielle, en particulier en ce qui concerne les températures extrêmes. (...) Certains épisodes récents de chaleur extrême auraient eu très peu de chances de se produire sans l'influence humaine sur le système climatique. (...) Même des augmentations relativement faibles du réchauffement climatique (+0,5°C) entraînent des changements statistiquement significatifs au niveau des phénomènes extrêmes à l'échelle mondiale et dans de vastes régions (degré de confiance élevé). (...) Outre l'augmentation de la température mondiale à la surface, de nombreux types de phénomènes météorologiques et climatiques extrêmes ont évolué. Dans la plupart des régions, la fréquence et l'intensité des vagues de chaleur ont augmenté, tandis que celles des vagues de froid ont diminué.

S'il est difficile d'identifier les causes exactes d'un événement extrême donné, la science relativement récente de l'attribution des événements permet de quantifier le rôle du changement climatique dans la modification de la probabilité et de l'ampleur de certains types de phénomènes météorologiques et climatiques extrêmes. (...) Au cas par cas, les scientifiques sont désormais en mesure de quantifier la contribution des activités humaines à l'ampleur et à la probabilité de nombreux phénomènes extrêmes. Pour ce faire, ils estiment et comparent la probabilité ou l'ampleur d'un même type de phénomène entre le climat actuel – compte tenu de l'augmentation des concentrations de gaz à effet de serre et d'autres influences humaines – et un monde alternatif où les gaz à effet de serre atmosphériques seraient restés à des niveaux pré-industriels.

¹Le chapitre 11 s'appelle *Weather and Climate Extreme Events in a Changing Climate*, et il est lié au chapitre 10 du 5ème rapport du GIEC qui s'appelait *Detection and Attribution of Climate Change : from Global to Regional*, et parlait plus largement de l'attribution de changements de type "medians".

La figure ci-contre, également extraite du chapitre 11 du 6^{ième} rapport du GIEC, illustre ce phénomène : un événement d'une certaine amplitude (une température donnée par exemple) devient plus fréquent dans le climat actuel, par rapport au climat pré-industriel, alors qu'un événement d'une probabilité donnée devient plus intense.

L'activité (de statistique) mathématique proposée ici s'inspire de l'article *Human contribution to the European heatwave of 2003* de Peter A. Stott, D. A. Stone et M. R. Allen, publié dans la revue *Nature* en 2004, un des premiers articles à présenter une étude d'attribution d'un événement extrême sur la base d'une approche qu'on qualifiera par la suite de probabiliste. L'objectif était d'évaluer la responsabilité humaine (à travers les émissions de gaz à effet de serre) dans la survenue de la vague de chaleur qui s'est abattue sur la France et une grande partie de l'Europe à l'été 2003.

Même si un phénomène météorologique de ce type aurait pu se produire dans un climat pré-industriel (du fait de la variabilité naturelle du climat), il est possible d'estimer dans quelle mesure les activités humaines ont pu accroître le risque de survenue d'une telle vague de chaleur.

Comme cela sera expliqué plus en détail plus loin, la démarche consiste à utiliser un modèle climatique pour simuler deux jeux de données : dans l'un on reproduit le climat tel qu'il existait lors de l'événement qui nous intéresse et dans l'autre, on reproduit le climat pré-industriel, en particulier sans émissions de gaz à effet de serre d'origine humaine. En comparant les deux, on peut mesurer à quel point le réchauffement a modifié la fréquence ou l'intensité d'un tel événement. Remarquons que lorsque l'étude ne s'intéresse pas à un événement particulier, on parle d'attribution d'un changement plutôt que d'attribution d'un événement. L'application numérique proposée dans la partie 2 s'inscrira dans ce cadre d'attribution d'un changement.

FAQ 11.3: Climate change and extreme events

Extreme events have become more probable and more intense. Many of these changes can be attributed to human influence on the climate.

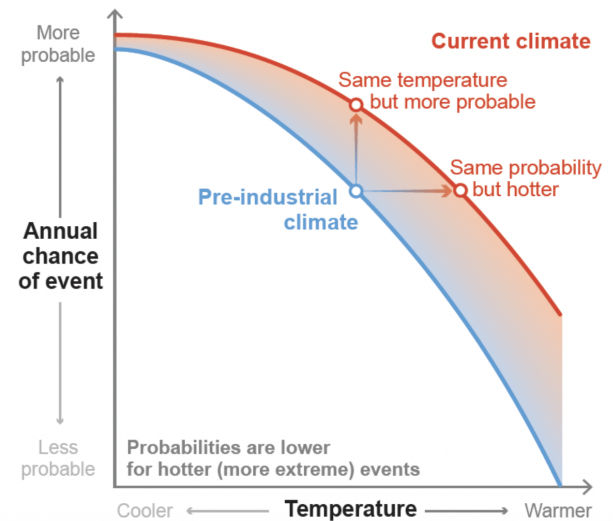


Figure 1: Illustration "Climate change and extreme events" tirée de 6^{ième} rapport du GIEC Chapitre 11 : Weather and Climate Extreme Events in a Changing Climate Cambridge University Press; p. 1611 [CC-BY-NC-ND 4.0]

1 Première partie : aspects théoriques

1.1 Variables et objet d'étude

Nous nous intéressons aux températures dans une région donnée, en cherchant à répondre à la question : *peut-on attribuer aux activités humaines la hausse du risque de températures extrêmes ?* Pour cela, nous supposons disposer de deux échantillons correspondant aux données de température, issues de modèles climatiques numériques² :

- ▷ le premier échantillon, $(X_1, \dots, X_n)$, est constitué de n variables aléatoires indépendantes et de même loi de fonction de répartition F , représentant les données *avec forçage naturel (solaire & aérosols dûs à l'activité volcanique) ET forçage anthropique (gaz à effet de serre supplémentaires dûs aux activités humaines)*. On parle alors de monde **historique** ou **factuel**.
- ▷ le second échantillon, $(Y_1, \dots, Y_n)$, est constitué de n variables aléatoires indépendantes et de même loi de fonction de répartition G , représentant les données *avec forçage naturel mais sans forçage radiatif anthropique*. On parle de monde **naturel** ou **contre-factuel**, correspondant à peu près à la période pré-industrielle.

Les deux échantillons sont supposés *indépendants*, et les variables X et Y seront dans ce TP **des maxima annuels de températures maximales journalières**, moyennisés dans une région d'intérêt³. Les différentes valeurs d'un échantillon correspondent aux différentes années successives où ce maximum annuel a été mesuré. Nous verrons dans la seconde partie comment ces données sont obtenues.

Le but du TP est d'estimer le **rapport de risque RR** (appelé aussi communément *risk-ratio*) défini comme le rapport entre la probabilité p_1 d'un événement donné dans un scénario factuel et la probabilité p_0 du même événement dans un scénario contre-factuel. **L'événement considéré sera le dépassement d'une température t relativement élevée**⁴ (qu'il faudra choisir à un moment). On a donc comme cible statistique (voir figure 2)

$$RR = \frac{p_1}{p_0} \quad \text{où} \quad p_1 = \mathbb{P}(X > t) \quad \text{et} \quad p_0 = \mathbb{P}(Y > t),$$

où l'on rappelle que X est de loi F (monde factuel) et Y de loi G (monde contre-factuel). Pour cela, il va falloir adopter un modèle statistique paramétrique pour les lois F et G , exprimer les quantités d'intérêt p_1 , p_0 et RR en fonction de ces paramètres, en construire des estimateurs et décrire leurs propriétés en termes de distributions approchées. En bout de course, il sera possible d'établir un intervalle de confiance pour RR , et ainsi en tirer des conclusions.

²Une question naturelle qui se pose est "Pourquoi utiliser uniquement des données de simulation ? Pourquoi ne pas utiliser des observations pour le monde factuel, et des données de simulation uniquement pour le monde contrefactuel ?". Il y a plusieurs réponses à cela. L'une d'elles est qu'il n'est pas si évident que cela de confronter des données d'observation à des données simulées, en raison des *biais* des modèles : il est plus sûr de comparer deux jeux de données venant d'un *même* modèle, le biais simulation-réalité ne pouvant alors pas brouiller la conclusion. Une autre raison est que les données d'observation sont souvent disponibles avec une résolution inhomogène, ce qui rendrait difficile les analyses de grande échelle.

³Du fait de cette moyennisation spatiale, les maxima pourront peut être paraître, au moment de l'application numérique, un peu en deçà de ce que tout un chacun vit à une échelle plus locale.

⁴Pour ne pas alourdir les notations, on ne fera en général pas apparaître la dépendance en t , dans les différentes expressions.

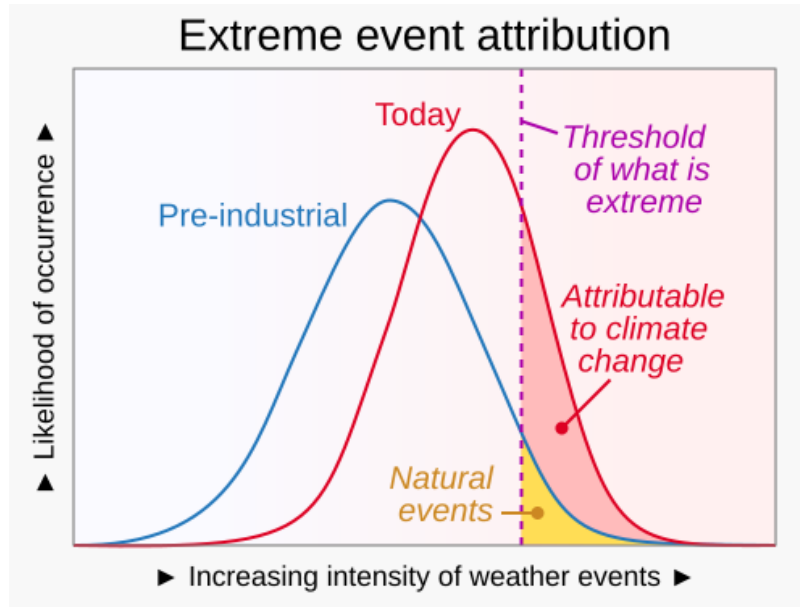


Figure 2: Illustration tirée de la page Wikipedia "Extreme Event Attribution" : à quoi correspondent le trait vertical pointillé et les aires rouges et jaunes ? [CC BY-SA 4.0]

Question 1 : Comment doit se situer $\hat{p}_1$ par rapport à $\hat{p}_0$ pour qu'il y ait attribution du risque de température extrême aux activités humaines, avec $\hat{p}_1$ et $\hat{p}_0$, respectivement estimateurs de p_1 et p_0 ?

Plus précisément, si l'on réussit à construire un *intervalle de confiance* pour RR , quelle valeur spécifique cherchera-t-on à situer par rapport à cet intervalle ? À quel test statistique cette démarche correspondra-t-elle ? (préciser les hypothèses H_0 et H_1)

Correction : Clairement, si les activités humaines ont tendance à rendre la survenue de hautes températures plus vraisemblable, alors on aura $p_1 > p_0$ pour la plupart des t (sinon tous les t), et donc on s'attend à observer $\hat{p}_1 > \hat{p}_0$. Le cœur de ce TP est d'être capable de quantifier le degré de confiance avec lequel on en déduira que $p_1 > p_0$.
Par conséquent, si on a un intervalle de confiance pour RR basé sur $\widehat{RR} = \hat{p}_1/\hat{p}_0$, on cherchera à savoir si la valeur 1 appartient à cet intervalle, ce qui reviendra à réaliser un test de $H_0 : "p_1 = p_0"$ contre $H_1 : "p_1 > p_0"$.

1.2 Méthode statistique

En vertu de la théorie des valeurs extrêmes ⁵, un modèle statistique naturel est ici de considérer que F et G appartiennent à la famille des *lois des Valeurs Extrêmes Généralisées (GEV)* de la forme

$$F(x) = H_{\gamma_X, \sigma_X, \mu_X}(x) \quad \text{et} \quad G(x) = H_{\gamma_Y, \sigma_Y, \mu_Y}(x),$$

⁵Le théorème de Fisher-Tippett stipule que la loi du maximum d'un grand nombre de variables aléatoires i.i.d. est approximable par une loi des Valeurs Extrêmes Généralisée ; voir par exemple le livre de S.Coles *An introduction to statistical modeling of extreme values* (Springer, 2001).

où

$$H_{\gamma,\sigma,\mu}(x) = \exp \left(- \left(1 + \frac{\gamma}{\sigma}(x - \mu) \right)_+^{-1/\gamma} \right)$$

avec $(f(x))_+ = \max\{f(x), 0\}$ et $(f(x))_+^a$ est une manière rapide d'écrire $((f(x))_+)^a$.

Dans une loi GEV, γ est le paramètre de *forme*, $\sigma > 0$ le paramètre d'*échelle* et μ le paramètre de *position*⁶. Notre modèle a donc au total 2 vecteurs de paramètres

$$\theta_X = (\gamma_X, \sigma_X, \mu_X) \quad \text{et} \quad \theta_Y = (\gamma_Y, \sigma_Y, \mu_Y)$$

qu'il va falloir estimer à l'aide des échantillons $(X_i)_{i \leq n}$ et $(Y_i)_{i \leq n}$.

Question 2 : En distinguant les cas où le paramètre de forme est négatif ou positif, déterminer le support des lois F et G , en introduisant la notation $\tau = \mu - \sigma/\gamma$.

Après avoir installé puis chargé le package `evd` de `R`, commenter ce que produit les commandes suivantes :

```
x=seq(-3,10,by=0.01)
plot(x,dgev(x,shape=0.2),col="red",type="l")
lines(x,dgev(x,shape=-0.1),col="blue")
lines(x,dgev(x,shape=0.6),col="black")
```

Faire d'autres tracés en faisant varier les autres paramètres de la GEV et les commenter.

En rassemblant les éléments précédents, et dans la mesure où le phénomène observé est une température, doit-on s'attendre à ce que les paramètres de forme γ_X et γ_Y soient plutôt positifs ou bien négatifs ?

Correction :

Si $\gamma > 0$, on a $1 + \frac{\gamma}{\sigma}(x - \mu) \geq 0 \Leftrightarrow x \geq \tau$, avec ici $H_{\gamma,\sigma,\mu}(\tau) = \exp(-0^{-1/\gamma}) = 0$, donc le support de la GEV est $[\tau, +\infty[$: τ est le point terminal gauche de la distribution, et la variable est non majorée.

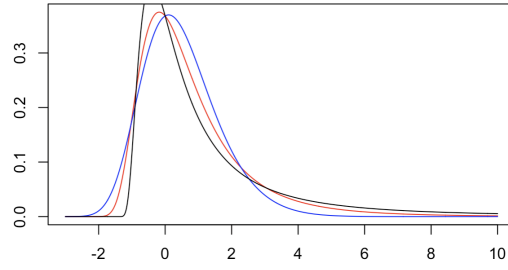
Si par contre $\gamma < 0$, on a $1 + \frac{\gamma}{\sigma}(x - \mu) \geq 0 \Leftrightarrow x \leq \tau$, avec ici $H_{\gamma,\sigma,\mu}(\tau) = \exp(-0^{-1/\gamma}) = 1$, donc le support de la GEV est $] -\infty, \tau]$: τ est ici le point terminal droit de la distribution, et la variable est majorée par τ .

Les 3 densités tracées à l'aide des commandes fournies illustrent le fait que plus γ est grand, plus la queue droite de la distribution s'épaissit, c'est-à-dire la variable de loi GEV a plus de chances de valoir de grandes valeurs. Cela s'accorde bien avec les considérations précédentes sur le support. On note bien que la densité bleue, avec $\gamma = -0.1 < 0$, devient nulle au delà de son point terminal droit, mais elle "traîne" un peu, car visuellement on a l'impression que ce point terminal droit est autour de 6, alors qu'il vaut $\mu - \sigma/\gamma = 0 - 1/(-0.1) = 10$!

On pourra aussi remarquer qu'il est insuffisant de comparer les paramètres de forme, les deux autres paramètres jouant aussi un rôle.

Enfin, si on analyse des données de températures, on s'attend bien sûr à ce que la distribution soit majorée, car les températures ne peuvent pas prendre des valeurs exagérément grandes, donc on s'attend à observer $\hat{\gamma} < 0$. On verra si ce sera le cas pour les données contre-factuelles (c'est-à-dire comparables à la situation pré-industrielle) et aussi pour les données factuelles.

⁶En anglais et dans les logiciels, on parle de *shape*, *scale* et *location parameters*. Une famille $\mathcal{F}$ de lois est dite (*shape*-)*location-scale* si toute densité $f \in \mathcal{F}$ s'écrit sous la forme $f(\cdot) = f_\gamma^* \left(\frac{\cdot - \mu}{\sigma} \right)$ pour une certaine densité $f_\gamma^* \in \mathcal{F}$ et certaines valeurs de μ et $\sigma > 0$. Dans le cas présent, la densité f_γ^* est celle de la loi GEV $H_{\gamma,1,0}$.



Question 3 : Déterminer l'expression de la fonction Φ telle que les quantités à estimer p_1 et p_0 s'expriment en fonction des paramètres par

$$p_1 = \Phi(\theta_X) \quad \text{et} \quad p_0 = \Phi(\theta_Y).$$

(Question optionnelle) En introduisant la fonction $h(\theta) = \left(1 + \frac{\gamma}{\sigma}(t - \mu)\right)^{-1/\gamma}$, vérifier que les dérivées partielles de la fonction h s'écrivent⁷

$$\begin{aligned} \frac{dh}{d\gamma}(\theta) &= h(\theta) \times \left(-\frac{1}{\gamma} \log(h(\theta)) - \frac{t-\mu}{\gamma\sigma} (h(\theta))^\gamma\right) \\ \frac{dh}{d\sigma}(\theta) &= \frac{t-\mu}{\sigma^2} (h(\theta))^{1+\gamma} \\ \frac{dh}{d\mu}(\theta) &= \frac{1}{\sigma} (h(\theta))^{1+\gamma} \end{aligned}$$

En déduire l'expression du gradient de la fonction Φ en fonction de celui de h et de l'expression $\Phi(\theta)$.

Correction : La première question est triviale : on a

$$p_1 = \mathbb{P}(X > t) = 1 - H_{\theta_X}(t) = 1 - \exp\left(-\left(1 + \frac{\gamma_X}{\sigma_X}(t - \mu_X)\right)_+^{-1/\gamma_X}\right) =: \Phi(\theta_X).$$

Pour la question optionnelle, on a $\Phi(\theta) = 1 - \exp(-h(\theta))$ avec (pour θ tel que t est $<$ ou $>$ à $\tau = \mu - \sigma/\gamma$, cf Question 2)

$$h(\theta) = \left(1 + \frac{\gamma}{\sigma}(x - \mu)\right)^{-1/\gamma} = \exp\left\{-\gamma^{-1} \log\left(1 + \gamma\sigma^{-1}(t - \mu)\right)\right\}.$$

En remarquant que $(1 + (\gamma/\sigma)(t - \mu))^{-1} = (h(\theta))^\gamma$, on aboutit après quelques efforts aux formules annoncées pour les dérivées partielles de h .

Quant à celles de Φ , elles s'en déduisent avec par exemple $\frac{d\Phi}{d\gamma}(\theta) = \frac{dh}{d\gamma}(\theta) \times \exp(-h(\theta))$. Cela procurera le gradient $\nabla\Phi(\theta)$ qui sera utile dans la question suivante.

Ainsi, pour estimer RR , il suffit d'estimer tous ces paramètres de façon paramétrique, par la méthode du maximum de vraisemblance par exemple ; le paramètre θ_X sera bien sûr estimé à partir de l'échantillon $(X_1, \dots, X_n)$ et θ_Y à partir de l'échantillon $(Y_1, \dots, Y_n)$.

Notons $\hat{\theta}_X$ et $\hat{\theta}_Y$ les estimateurs du maximum de vraisemblance (EMV) de θ_X et θ_Y respectivement. Nous admettons que le modèle GEV est *régulier* lorsque $\gamma_X > -1/2$ et $\gamma_Y > -1/2$. Nous disposons alors dans ce cas des résultats asymptotiques suivants, qui ont peut-être déjà été vus en cours de statistique mathématique : quand n tend vers l'infini,

$$\begin{aligned} \sqrt{n}(\hat{\theta}_X - \theta_X) &\xrightarrow{\text{loi}} \mathcal{N}_3(0, I^{-1}(\theta_X)), \\ \sqrt{n}(\hat{\theta}_Y - \theta_Y) &\xrightarrow{\text{loi}} \mathcal{N}_3(0, I^{-1}(\theta_Y)), \end{aligned}$$

où $I^{-1}(\theta)$ désigne l'inverse de la matrice d'information de Fisher pour le modèle GEV, et $\mathcal{N}_3$ une loi normale multivariée de dimension 3. Les expressions des différents termes de la matrice $I(\theta)$ pour la famille des lois GEV ne sont pas faciles à obtenir, mais sont fournies en annexe A à titre indicatif.

Question 4 : En utilisant la delta-méthode (cf annexe B), identifier les variances $v_1(\theta_X)$ et $v_0(\theta_Y)$ dans les résultats asymptotiques suivants, en fonction des notations introduites jusque là :

$$\begin{aligned}\sqrt{n}(\hat{p}_1 - p_1) &\xrightarrow{loi} \mathcal{N}(0, v_1(\theta_X)), \\ \sqrt{n}(\hat{p}_0 - p_0) &\xrightarrow{loi} \mathcal{N}(0, v_0(\theta_Y)).\end{aligned}$$

Par l'indépendance supposée entre les échantillons (X_i) et (Y_i) , on en déduit que

$$\sqrt{n} \begin{pmatrix} \hat{p}_1 - p_1 \\ \hat{p}_0 - p_0 \end{pmatrix} \xrightarrow{loi} \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} v_1(\theta_X) & 0 \\ 0 & v_0(\theta_Y) \end{pmatrix} \right).$$

En utilisant alors encore la delta-méthode, en déduire le résultat asymptotique final pour la loi de $\log \widehat{RR} - \log RR$ où $\widehat{RR} = \hat{p}_1/\hat{p}_0$ est l'estimateur naturel de RR .

Correction :

En se basant sur la Question 3, on a $\hat{p}_1 = \Phi(\hat{\theta}_X)$ et $p_1 = \Phi(\theta_X)$, donc une application directe de la delta-méthode entraîne que $v_1(\theta_X) = (\nabla \Phi(\theta_X))^t \cdot I^{-1}(\theta_X) \cdot \nabla \Phi(\theta_X)$.

(ici, l'exposant t désigne la transposée, sans lien avec le seuil t considéré dans cette activité).

Pour information, la convergence en loi bivariée annoncée ci-dessus provient des propriétés des vecteurs gaussiens (cela se prouve via les fonctions caractéristiques, en utilisant l'indépendance des variables $\hat{p}_1$ et $\hat{p}_0$ elle-même issue de celle des échantillons X et Y). On peut alors s'atteler à la normalité asymptotique de

$$\log \widehat{RR} = \log \frac{\hat{p}_1}{\hat{p}_0} = g(\hat{p}_1, \hat{p}_0)$$

où $g(x, y) = \log(x) - \log(y)$, donc $\nabla g(p_1, p_0) = (1/p_1, -1/p_0)$, et la delta-méthode implique que

$$\sqrt{n} \log(\widehat{RR}/RR) \xrightarrow{loi} \mathcal{N} \left(0, \frac{v_1(\theta_X)}{p_1^2} + \frac{v_0(\theta_Y)}{p_0^2} \right).$$

Question 5 : En déduire un intervalle de confiance pour RR de niveau approché $1 - \alpha$, pour α donné dans $]0, 1[$, en introduisant les notations :

$$\sigma_{RR}^2 := p_1^{-2} v_1(\theta_X) + p_0^{-2} v_0(\theta_Y) \quad \text{et} \quad \hat{\sigma}_{RR}^2 := \hat{p}_1^{-2} v_1(\hat{\theta}_X) + \hat{p}_0^{-2} v_0(\hat{\theta}_Y).$$

Correction :

Par la consistance des EMV $\hat{\theta}_X$ et $\hat{\theta}_Y$, on a celle de $\hat{p}_1$ et $\hat{p}_0$ et ainsi le Lemme de Slutsky + le résultat de la question précédente implique que $\sqrt{n} \log(\widehat{RR}/RR)/\hat{\sigma}_{RR}$ converge en loi vers une gaussienne centrée réduite. Cela va être notre statistique pivotale pour déduire un intervalle de confiance pour RR .

En effet, pour $\alpha \in]0, 1[$ donné, si $\mathbb{P}(\mathcal{N}(0, 1) > z) = \alpha/2$, on a, pour n grand,

$$1 - \alpha = \mathbb{P}(-z \leq \mathcal{N}(0, 1) \leq z) \simeq \mathbb{P}\left(\exp(-z\hat{\sigma}_{RR}/\sqrt{n}) \leq \widehat{RR}/RR \leq \exp(z\hat{\sigma}_{RR}/\sqrt{n})\right)$$

et par conséquent l'intervalle de confiance recherché pour RR , de niveau approché $1 - \alpha$, est $[\widehat{RR} \times e^{-W}, \widehat{RR} \times e^W]$ où $W := z\hat{\sigma}_{RR}/\sqrt{n}$. Plus n sera grand et/ou la variance estimée sera faible, plus le facteur e^{-W} sera proche de 1 et donc l'intervalle de confiance sera étroit. On remarque que plus le seuil choisi t sera grand, plus les probabilités p_1 et p_0 seront faibles et plus la variance sera élevée.

Attention à une confusion qui va probablement survenir dans l'application numérique : la standard-error de $\log \widehat{RR}$ est l'estimation de l'écart-type de $\log \widehat{RR}$, et est donc égal à $\hat{\sigma}_{RR}/\sqrt{n}$, racine de n compris !

Dans la suite, une fonction prédéfinie dans le logiciel R sera utilisée pour estimer ces paramètres par maximum de vraisemblance et pour déterminer la valeur observée de l'intervalle de confiance pour RR ; les étudiant·e·s intéressé·e·s pourront essayer de vérifier que les formules obtenues plus haut mènent aux mêmes résultats.

2 Seconde partie : application numérique

Cette seconde partie est consacrée à une application de la procédure construite précédemment : rappelons que nous n'allons pas considérer l'attribution d'un événement extrême particulier (pour lequel il est difficile d'avoir accès à des données exploitables), mais plutôt considérer l'*attribution d'un changement* (dans l'ampleur des phénomènes extrêmes).

Chaque groupe d'étudiant·es est chargé d'un jeu de données qui lui est propre. Il s'agit de données de maxima annuels de températures (le jargon est *tasmax* = *daily maximum of near-surface air temperature*) pour une région française donnée (température moyennisée sur toute la zone).

Le modèle dont sont issues ces données est le modèle océan-atmosphère CERFACS.CNRM-CM6-1-HR, version dite "à haute résolution" du modèle climatique CM6-1 de Météo-France. Pour ce modèle, comme dans la plupart des modèles régionaux, il n'existe pas de données factuelles et contrefactuelles correspondant aux échantillons X et Y tels que décrits dans la section 1.1. À la place, nous disposons de données issues d'un modèle factuel (correspondant aux conditions climatiques réelles) sur une large période (de 1850 à 2100).

Les données fournies⁸ dans cet atelier, sont extraites de ces simulations . Il s'agit de simulations historiques (monde factuel) jusqu'en 2014, et du scénario⁹ SSP5-8.5 du GIEC après. Toutefois,

⁸Merci à Soulivanh Thao (membre du Laboratoire des Sciences du Climat et de l'Environnement), pour avoir aidé à accéder à ces données issues de PCMDI: Earth System Model Evaluation Project, notamment CMIP6 model data (CMIP6 - Coupled Model Intercomparison Project Phase 6) [CC BY 4.0]

⁹Voir Shared Socioeconomic Pathways pour une explication de ce que sont les scénarios SSP utilisés pendant l'essentiel des années 2020.

les trente¹⁰ premières années (1850 – 1880) peuvent être considérées comme un échantillon contre-factuel Y car les émissions de gaz à effet de serre (anthropogéniques) ne sont pas encore trop présentes à cette période. On prendra comme échantillon X factuel les données correspondant aux années 2000 – 2030 ou 2020 – 2050.

Comme on s'intéresse, dans cette partie, à l'attribution d'un changement, il n'est pas trop gênant de ne pas avoir de données factuelles et contre-factuelles correspondant à une même période. De plus, rappelons qu'il s'agit ici d'une illustration à but pédagogique : dans la pratique, une plus grande taille d'échantillons (ou bien plusieurs séries de simulations) est nécessaire.

2.1 Analyse des deux échantillons factuels et contre factuels

Question 6 : Après avoir récupéré les données correspondant à sa région avec la commande `read.csv("yearlymax_tasmax.csv", sep=";", dec=",")`, il faut prélever les deux échantillons X et Y et les comparer à l'aide de la fonction `summary` du logiciel R.

Correction : On constate simplement que l'échantillon factuel est "décalé" vers la droite par rapport à l'échantillon contre-factuel, en termes de moyenne, médiane, troisième quartile et maximum. Par exemple, pour la région Île-de-France, on obtient

Échantillon	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Factuel (X)	31.27	34.76	35.88	35.75	36.86	39.10
Contre-Factuel (Y)	30.25	32.36	34.52	33.90	35.41	38.34

La question suivante permettra de visualiser les choses, tout en constatant qu'une taille d'échantillons de 31 n'est pas optimale pour faire de la statistique.

Question 7 : En utilisant la fonction `hist`, tracer les histogrammes des deux échantillons X et Y sur l'intervalle $[25, 45]$ (utiliser l'option `xlim=c(25,45)`). Qu'observe-t-on ? Faire le lien avec la question précédente et la figure 2.

(les plus connaisseurs en R pourront obtenir les 2 histogrammes l'un sur l'autre dans le même graphique avec `ggplot()`, en rassemblant les 2 échantillons dans un même data-frame dont la première colonne contient les 2 échantillons empilés, et la deuxième contient la variable catégorielle précisant l'échantillon associé)

Correction : Voilà ce que cela donne avec `hist` pour les données de l'Île-de-France :

```
hist(X,xlim=c(25,45),col="red")
hist(Y,xlim=c(25,45),col="green")
```

Avec `ggplot`, voici un code possible et le résultat en figure (3) :

```
n <- nrow(dataX)
don <- data.frame(c(donX,donY),rep(c("F","CF"),c(n,n)))
names(don) <- c("tasmax","monde")
ggplot(don,aes(x=tasmax,col=monde)) +
  geom_histogram(fill="white",alpha=0.5,position="identity")
```

Cette question est surtout le prétexte pour visualiser les données, et rappeler combien il est important de le faire. On fait en tout cas le même constat de décalage que dans la question précédente.

¹⁰On se restreint à 30 ans afin de pouvoir considérer les membre $X_1, \dots, X_n$ ou $Y_1, \dots, Y_n$ comme étant de même loi, le régime climatique étant supposé stationnaire sur cette durée.

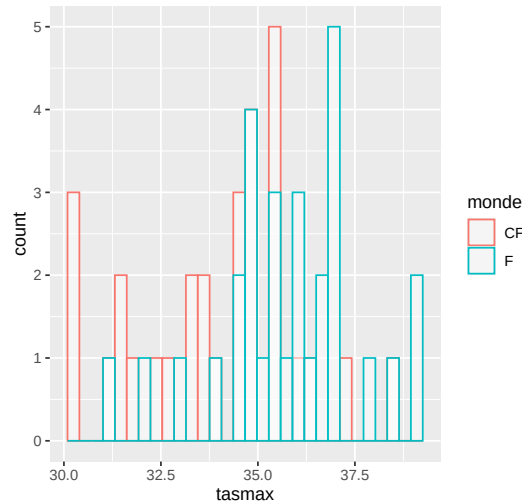


Figure 3: Histogrammes des échantillons X et Y pour l'Île-de-France

2.2 Estimation des différents paramètres

Question 8 : Utiliser la fonction `fit_gev` du package "climextRemes" pour calculer des intervalles de confiance de niveau 95% pour les paramètres γ_X et γ_Y . Comparer les valeurs obtenues et tracer les deux densités estimées. Que peut-on observer ?

Correction : On doit en théorie avoir $\hat{\gamma}_Y < 0$ et $\hat{\gamma}_X < 0$ et un intervalle de confiance pour γ_X est donné par $[\hat{\gamma}_X - 1.96se, \hat{\gamma}_X + 1.96se]$ (même chose pour Y), où se est donné en sortie de la fonction. On peut remarquer que γ_X n'est pas forcément plus grand que γ_Y car la loi ne dépend pas uniquement du paramètre de forme. Par contre, on observe bien un "décalage" vers la droite de la densité estimée de X par rapport à celle de Y , ce qui est en accord avec la figure 2.

Question 9 : Pour le seuil t à choisir, utiliser la fonction `calc_riskRatio_gev` du package "climextRemes" pour obtenir un intervalle de confiance pour RR de niveau 95%. Que peut-on en conclure (cf question 1) ?

Correction : Ici on regardera bien sûr si 1 appartient à l'intervalle de confiance observé. On peut même regarder si 2 ou 3 appartiennent à l'intervalle, ce qui voudrait dire qu'on a une confiance suffisante pour affirmer que le risque d'expérimenter l'événement extrême est 2 ou 3 fois plus élevé avec le forçage anthropique que dans la situation pré-industrielle. La confrontation des résultats des différents groupes d'étudiant-e-s est très encouragée.

Les étudiant-e-s les plus courageux pourront chez eux/elles mettre en œuvre les calculs de la partie théorique pour voir si cela correspond bien aux sorties "automatiques" de la fonction `calc_riskRatio_gev`.

Question 10 : Reprendre la question 9 avec un seuil t' plus grand que t . Commenter ce que vous observez.

Correction :

Le but ici est de faire constater que quand on prend un événement plus extrême, avec t' plus grand que t , on aboutit certes à des probabilités p_1 et p_0 plus faibles, mais aussi à une plus grande incertitude (intervalle de confiance plus large, ne permettant en général pas de conclure fermement à une attribution) et il peut être intéressant de voir comment le risk-ratio évolue quand t est de plus en plus grand.

Cette question peut être l'occasion de discussion dans le groupe classe sur l'influence de la faible valeur $n = 30$ sur la force de l'inférence (la valeur $n = 30$ étant due à la contrainte de stationarité imposée à l'échantillon X), et sur le choix du seuil t (et de manière générale sur "comment fait-on pour comparer 2 lois sans se limiter à comparer simplement une seule de leurs caractéristiques"). Ou encore sur ce qu'il faudrait modifier si l'on prenait des tailles différentes pour les échantillons X et Y (en prenant le ratio des tailles équivalent à une constante, il est possible de modifier les résultats des questions 4 et 5 en conséquence).

Conclusion

La science de l'attribution est une science encore jeune, mais qui progresse rapidement. Depuis quelques années, différentes équipes dans le monde (World Weather Attribution, Climameter, Copernicus) cherchent à analyser de plus en plus systématiquement les événements extrêmes dès leur survenue et à publier leurs résultats peu de temps après. C'est ce qui est parfois appelé les *rapid attribution studies*.

Voici quelques exemples où cette science en plein essor peut jouer un rôle :

- Aider les décideurs (et le public) à comprendre la responsabilité humaine.
- Aider à élaborer des mesures et politiques d'adaptation au changement climatique en montrant que certains événements deviennent plus fréquents et plus intenses.
- Pointer les responsabilités pour justifier les compensations financières dans les négociations internationales.
- Servir de base scientifique aux actions en justice pour réparation des dommages.¹¹

Un petit bémol tout de même au sujet de l'utilisation croissante de l'IA dans ces études d'attribution et du fait que ces travaux demandent de grosses capacités de calculs. Est-ce-qu'on ne participe pas au problème en voulant le résoudre ?

¹¹Voir cet article de Courier International

Annexe A : information de Fisher pour la loi GEV

Soit I la matrice 3×3 d'information de Fisher associée au modèle $H_{\gamma, \sigma, \mu}$ (de densité $h = h_{\gamma, \sigma, \mu}$) :

$$I = -\mathbb{E}_{\theta} \left[\text{Hess}_{\log h(X, \cdot)}(\theta) \right] = \left(-\mathbb{E}_{\theta} \left[\left(\frac{\partial^2}{\partial \theta_j \partial \theta_k} \log h(X, \theta) \right) \right] \right)_{1 \leq j \leq 3, 1 \leq k \leq 3}.$$

D'après l'article *Maximum Likelihood Estimation of the Parameters of the Generalized Extreme-Value Distribution* de Prescott et Walden (Biometrika 67(3), 1980), les coefficients de I sont :

$$\begin{aligned} I_{1,1} &= \frac{1}{\gamma^2} \left(\frac{\pi^2}{6} + (1 - \eta + \frac{1}{\gamma})^2 - \frac{2q}{\gamma} + \frac{p}{\gamma^2} \right) \\ I_{1,2} = I_{2,1} &= \frac{1}{\sigma \gamma^2} \left(\eta + q - 1 - \frac{p}{\gamma} - \frac{1 - \Gamma(2 + \gamma)}{\gamma} \right) \\ I_{1,3} = I_{3,1} &= \frac{1}{\sigma \gamma} \left(\frac{p}{\gamma} - q \right) \\ I_{2,2} &= \frac{1}{\sigma^2 \gamma^2} (1 - 2\Gamma(2 + \gamma) + p) \\ I_{2,3} = I_{3,2} &= \frac{1}{\sigma^2 \gamma} (\Gamma(2 + \gamma) - p) \\ I_{3,3} &= \frac{p}{\sigma^2} \end{aligned}$$

où η désigne la constante d'Euler¹², $p = (1 + \gamma)^2 \Gamma(1 + 2\gamma)$ et $q = \Gamma(2 + \gamma) \left(\psi(1 + \gamma) + \frac{1 + \gamma}{\gamma} \right)$, et enfin ψ est la fonction dérivée du logarithme de la fonction Γ (appelée fonction *digamma*).

Annexe B : delta-méthode multivariée

La delta-méthode est un résultat probabiliste important, majoritairement utilisé en Statistique et basé sur la formule de Taylor, qui permet de transférer un résultat de normalité asymptotique d'une suite aléatoire à un résultat de normalité asymptotique pour une fonction non-linéaire¹³ de cette suite.

L'énoncé est le suivant. Soit $(Z_n)_{n \geq 1}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^p$, z un vecteur de $\mathbb{R}^p$ et g une fonction de $\mathbb{R}^p$ dans $\mathbb{R}$ différentiable en z de gradient $\nabla g(\cdot)$ non nul en z . Si

$$\sqrt{n}(Z_n - z) \xrightarrow{\text{loi}} \mathcal{N}_p(0, \Sigma),$$

pour une certaine matrice de covariance Σ , alors

$$\sqrt{n}(g(Z_n) - g(z)) \xrightarrow{\text{loi}} \mathcal{N}(0, (\nabla g(z))^t \cdot \Sigma \cdot \nabla g(z)).$$

¹²Usuellement la constante d'Euler est désignée par la lettre γ , mais celle-ci est réservée à un autre usage dans cette application.

¹³Si la transformation est linéaire, les propriétés des vecteurs gaussiens rendent la delta-méthode inutile.

Annexe C : Commandes pour l'utilisation du logiciel R

Pour ceux qui ne sont pas familiers avec le logiciel R, voici les commandes qui seront utiles dans le cadre de ce TP :

- Installation puis chargement des packages "evd" et "climextRemes" avec les commandes `install.packages("...")` puis `library("...")`.
- Création d'un vecteur avec `seq(min,max,by=pas)`, par exemple pour contenir les abscisses en lesquelles une fonction sera calculée, en vue de tracer son graphe.
- Tracé d'une fonction avec `plot(...,type="l")` (et `lines` pour en ajouter d'autres sur le même graphique) ou d'un histogramme avec `hist`. L'utilisation de `ggplot()` du package `ggplot2` est aussi possible.
- Autres fonctions utiles :
 - `summary` donne les caractéristiques centrales et de dispersion d'un échantillon
 - `dgev` calcule la densité d'une loi GEV
 - `fit_gev` pour calculer les estimations par maximum de vraisemblance des paramètres de la GEV
 - `calc_riskRatio_gev` pour estimer le risque RR pour un seuil donné et intervalle de confiance associé (dans le modèle GEV)
 - `names` pour connaître les noms des variables des champs des sorties d'une fonction (les sorties de fonction sont souvent des objets de type liste)
 - `help("nom de fonction")` pour obtenir de l'aide sur une fonction.

Ce TD a été mutualisé dans le cadre d'un partenariat entre l'Université Paris-Est-Créteil, l'Université de Versailles Saint-Quentin-en-Yvelines et la Fondation UVED pour la déclinaison disciplinaire de l'enseignement des enjeux de transition écologique et sociétale (TEDS).

Il est mis à disposition selon les termes de la [Licence Creative Commons - 4.0 International : Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions](#)



Pour la formation continue ou professionnelle, les modalités d'usage sont à déterminer avec UVED et doivent faire l'objet d'un contrat définissant les conditions d'usage et de commercialisation. Contact : contact@fondation-ued.fr

Première édition : octobre 2026