



Examen UE7 Statistiques

Automne 2021

M. Bailly-Bechet

Université Nice Côte d'Azur – France

Durée : 3h. Tous documents autorisés, communications interdites. Les parties sont indépendantes et peuvent être traitées dans l'ordre de votre choix. Votre rendu comportera deux parties : une copie papier qui comportera toutes les réponses rédigées, et un fichier PDF (issu d'un document Word ou LibreOffice) qui contiendra les commandes , les sorties  et les graphiques que vous voudrez ajouter à votre analyse. Ce document doit être déposé sur la boîte de dépôt Moodle du module.

Feux de forêts

Le jeu de données que vous allez analyser ici porte sur des feux de forêts, dans deux régions différentes fortement touchées en été. Chaque ligne correspond à une journée d'observation. Les colonnes correspondent à :

Region : "A" ou "B" pour indiquer les deux régions étudiées

Temperature : température à midi en degrés Celsius

RH : humidité relative en %

Ws : vitesse du vent en km/h

Rain : quantité totale de pluie sur la journée, en mm

FFMC : Fine Fuel Moisture Code index

DMC : Duff Moisture Code index

DC : Drought Code index

ISI : Initial Spread Index

BUI : Buildup Index

FWI : Fire Weather Index

Observation : 0 quand aucun incendie n'a été ou "not fire" selon qu'un feu a été observé dans la région le jour concerné

Les index FFMC, DMC, DC, ISI, BUI, et FWI sont des valeurs calculées par les services forestiers pour prévoir le risque d'incendie. Le FWI est la mesure finale employée, les autres servant de calculs intermédiaires.

Analyse exploratoire multivariée

Q1– Réalisez une analyse en composantes principales de ce jeu de données, sans prendre en compte les variables qualitatives **Observation** et **Region**, en conservant 2 axes. Quel est le pourcentage de variance expliquée par chacun des axes ?

Q2– Tracez le cercle de corrélation des différentes variables et commentez-le. Quelles variables semblent fortement corrélées ? Anticorrélées ? En particulier, que pouvez-vous dire des corrélations entre les variables **Rain** et **Temperature** et de leurs projections sur le plan factoriel à deux dimensions ?

Modélisation du Fire Weather Index (FWI)

On cherche dans cette partie à comprendre ce que représente le `FWI`, cette mesure intégrée qui sert au final de prédicteur pour les agents forestiers.

Q3– Pour vérifier si cette mesure est pertinente, commencez par tracer le graphique du **FWI** en fonction du fait qu’il y ait eu ou non un feu. Faîtes ensuite le test statistique approprié pour savoir si les mesures de **FWI** ont des moyennes différentes quand un feu a eu lieu ou non.

On dit souvent qu’une température élevée et un vent fort sont de bons prédicteurs d’incendie. On cherche ici à prédire avec une régression linéaire multiple le **FWI** à partir des variables **Temperature** et **Ws**.

Q4– Effectuez la régression linéaire multiple correspondante, sans interaction, et interprétez son résultat ; mentionnez en particulier quel est l’effet précis de chacun des facteurs significatifs.

Q5– Effectuez la même analyse avec interaction. Interprétez les effets précis de chaque variable et commentez les changements par rapport à la question précédente.

Q6– Les valeurs **FFMC**, **DMC**, **DC**, **ISI** et **BUI** sont des valeurs qui sont employées dans le calcul du **FWI**. On cherche à retrouver une équation simple, linéaire, permettant d’approximer le **FWI** à partir de ces 5 variables ou moins. Proposez un tel modèle et l’équation correspondante ; la démarche que vous emploieriez pour les trouver est aussi importante que le résultat final (par simplicité, on raisonnera sans interactions).

Q7– Tracez un graphique représentant les valeurs prédites par votre modèle en fonction des valeurs réelles de **FWI**, ainsi que la droite sur laquelle devraient se trouver les points si vos prédictions étaient parfaites. Si vous observez un ou des outliers, pouvez-vous émettre une hypothèse sur la raison de votre mauvaise prédiction pour ces points ? (*Si vous n’avez pas répondu à la question précédente, vous pouvez répondre à celle-ci en employant comme référence le modèle `lm(FWI ~ ISI+DC, data=forest)` par exemple.*)

Prédiction des feux de forêts

On cherche maintenant à prédire les feux de forêts observés (variable **Observation**) à l'aide uniquement des variables météorologiques directes et de la variable **Region**. On ne prend pas en compte la variable **Rain** car son effet serait trop important et "écraserait" les autres.

Q8– Effectuez et interprétez la régression logistique permettant de prédire les feux de forêts observés à l'aide simultanément des variables **Temperature**, **RH**, **Ws** et **Region**, sans interactions. Interprétez vos résultats : quels variables ont un effet significatif sur les feux de forêts ? En augmentent-elles la probabilité ou la diminuent-elles ?



Q9– Effectuez maintenant, et interprétez, la régression logistique permettant de faire cette prédiction avec uniquement la variable **Region** comme prédicteur. Dans quelle région prédisiez-vous le plus de feux ? Quelle serait la probabilité d'avoir un feu dans la région la plus touchée si la probabilité d'en avoir un dans la région la moins touchée était de 0.5, d'après votre modèle ?

Q10– Comparez les résultats des deux modèles précédents. Quelles hypothèses simples vous permettraient de rendre cohérents leurs résultats ? Comment pourriez-vous les vérifier ?

Q11– Il existe un autre test statistique permettant de traiter précisément des données présentées sous cette forme – **Region** et **Observation**, et uniquement ces deux variables – sans recourir au modèle linéaire ni à la régression logistique. Indiquez lequel, donnez ses hypothèses et appliquez-le sur ces données : que concluez-vous ?

Q12– ★ **Bonus, à ne faire qu'après tout le reste** On pourrait penser qu'employer les données du jour même pour prédire les feux de forêts est réducteur, et qu'il peut y avoir des effets cumulatifs – par exemple de fortes températures plusieurs jours de suite. En ne travaillant que sur une région, proposez une méthode de votre choix pour prédire les feux à partir des données de la veille (donc la ligne précédente), ou de la combinaison des données de la veille et du jour-même.¹

Bébés australiens

Dans cet exercice il vous est demandé d'interpréter des résultats  sur des tests statistiques que vous ne connaissez pas forcément. Il ne vous est pas nécessaire d'aller voir la documentation, vous pouvez répondre en vous basant uniquement sur les informations données dans l'énoncé et vos connaissances statistiques ; aucune manipulation  n'est nécessaire ici, même si vous pouvez récupérer le jeu de données si vous voulez en faire.

1. Oui, ça a l'air compliqué, et ça sort un peu du cadre du cours tel qu'on l'a vu. En même temps c'est un bonus si vous avez fait tout le reste, il faut bien que je vous occupe avant vos vacances !

On travaille sur un jeu de données appelé **babyboom** du package **UsingR** contenant des informations sur 44 bébés nés dans un hôpital de Brisbane (Etat du Queensland, Australie) sur une période de 24 heures :

clock.time l'heure affichée sur l'horloge au moment de la naissance, au format HHMM.

gender le sexe biologique : boy ou girl)

wt le poids de naissance exprimé en grammes

running.time le délai entre minuit et la naissance, exprimé en minutes (pour avoir une variable temporelle qui soit un vrai nombre, contrairement à **clock.time**)

```
> library(UsingR)
> data(babyboom)
> head(babyboom)

  clock.time gender   wt running.time
1          5  girl 3837             5
2         104  girl 3334            64
3         118   boy 3554            78
4         155   boy 3838           115
5         257   boy 3625           177
6         405  girl 2208           245

> dim(babyboom)
[1] 44  4
```

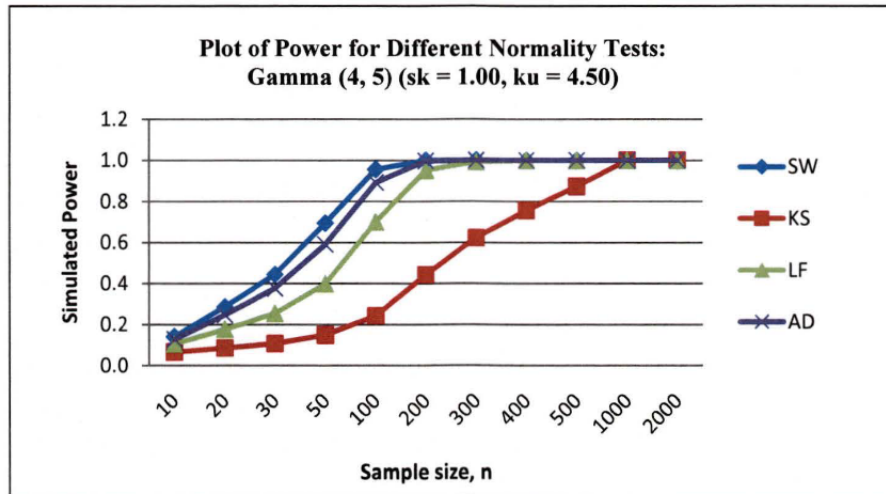
Test de Kolmogorov-Smirnov, Shapiro-Wilks et analyse de puissance

On veut tester la normalité du poids des 44 bébés. On va à des fins de comparaison utiliser deux tests différents : le test de Kolmogorov-Smirnov (un test qui permet de comparer la distribution de nos données à n'importe quelle distribution de référence choisie, ici une distribution normale en prenant comme référence une distribution normale ayant les même moyenne et écart-type que l'échantillon) et celui de Shapiro-Wilks, qui est un test spécifique pour vérifier la normalité. Les deux tests utilisent des statistiques différentes, mais ont les mêmes hypothèses :

H_0 : la variable étudiée est distribuée suivant une loi normale ;

H_1 : la variable étudiée n'est pas distribuée suivant une loi normale.

Avant de les employer sur nos données, une recherche bibliographique nous permet de les comparer. Dans un article publié en 2010 dans *Journal of Statistical Modeling and Analytics*, N. Razali et al. comparent par une approche de simulation la puissance des tests de Shapiro-Wilks et Kolmogorov-Smirnov (et deux autres tests) pour tester l'hypothèse nulle de normalité. Ils obtiennent le graphique suivant :



Axe des x : taille d'échantillon testée. Axe des y : Puissance mesurée. Le risque α était fixé à 0.05 pour toutes les simulations. Quatre tests vérifiant la normalité sont évalués : Shapiro-Wilks (SW), Kolmogorov-Smirnov (KS), Lielliefors (LF) et Anderson-Darling (AD). La taille d'effet, i.e le choix particulier de H_1 , est le même pour tous.²

Q13– Commentez le graphique ci-dessus en essayant d'en tirer les informations les plus pertinentes pour votre situation. En particulier, quelles sont, approximativement, les puissances des tests de Kolmogorov-Smirnov et de Shapiro-Wilks pour votre taille d'échantillon ?

On applique nos deux tests au poids des bébés. On obtient :

```
> shapiro.test(babyboom$wt)
      Shapiro-Wilk normality test

data:  babyboom$wt
W = 0.89872, p-value = 0.0009944

> ks.test(babyboom$wt,pnorm,mean(babyboom$wt),sd(babyboom$wt))
      One-sample Kolmogorov-Smirnov test

data:  babyboom$wt
D = 0.18336, p-value = 0.1038
alternative hypothesis: two-sided
```

Q14– Les deux tests précédents parviennent-ils à la même conclusion ? Pouvez-vous expliquer pourquoi ? Concluez sur la normalité des poids des bébés du Queensland à partir de ces tests.

2. Oui, de manière regrettable, cette figure de stats, publiée dans un journal de stats, a été produite avec Excel... mais c'était ça ou l'article original de Shapiro de 1965 ! Et bon, continuer à lire cette note de bas de page vous amusera peut-être³, mais ne vous fera pas avancer sur l'examen. Au boulot !

3. Ou pas. Mais j'aime les notes de bas de page récursives.

Une autre figure de l'article de Razali et al. montre que, dans un autre contexte, le test de Shapiro-Wilks a une puissance de 20%, pour un risque α de 5%. Cette figure est reprise sur un blog, qui en conclut que si résultat d'un test de Shapiro-Wilks est H_1 (l'hypothèse de non normalité) sur un jeu de données particulier, alors il n'y a que 20% de chances que cette hypothèse soit effectivement correcte.

Q15– L'interprétation ci-dessus est-elle correcte? Si oui, précisez-là en indiquant ce qu'on peut dire de plus au vu des données du problème. Si non, corrigez-là.

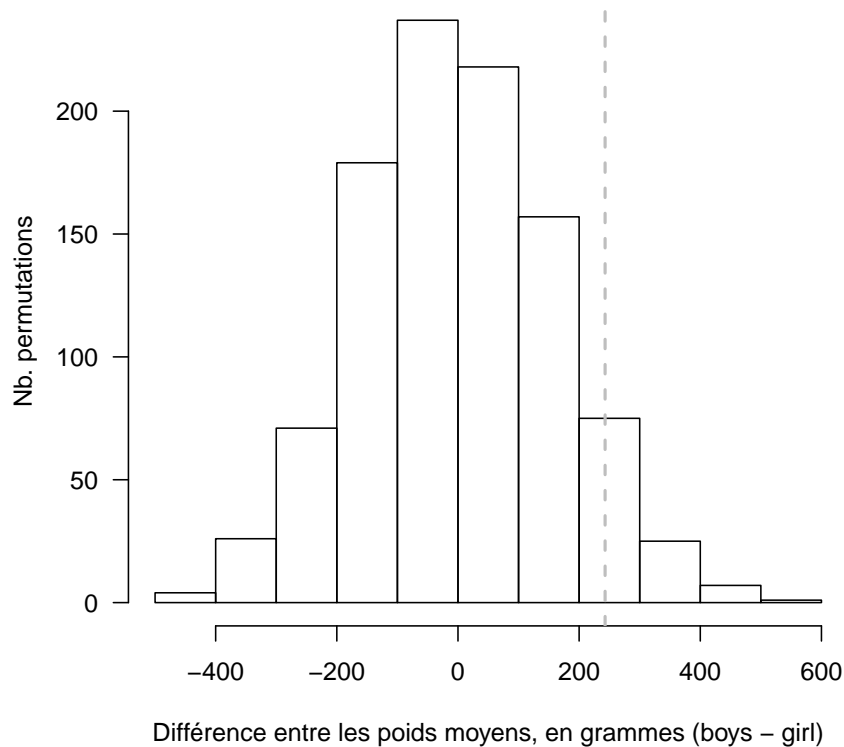
Approche par permutation

Pour déterminer si le poids des bébés garçons est plus important que le poids des bébés filles dans le Queensland, en se basant sur cet échantillon, un physicien veut faire un test statistique pour comparer leurs poids moyens.

Comme il n'a jamais suivi de cours de statistiques et ne sait pas quel test employer⁴, il applique une approche par permutation qui réattribue les labels "boy" et "girl" aux 44 poids de l'échantillon et calcule la différence entre les moyennes des poids des garçons et des filles. Sur la figure page suivante, il a tracé la distribution de ces valeurs (histogramme en noir) et y a superposé la valeur observée pour les données réelles (ligne pointillée grise). Finalement il calcule une p-valeur empirique. Le code est donné ci-dessous :


```
> weight_boys<- babyboom$wt[babyboom$gender=="boy"]
> weight_girls<- babyboom$wt[babyboom$gender=="girl"]
> real_delta<- mean(weight_boys)- mean(weight_girls)
> delta_means<-numeric(1000)
> for(i in 1:1000){
+   g<- sample(babyboom$gender)
+   weight_boys<- babyboom$wt[g=="boy"]
+   weight_girls<- babyboom$wt[g=="girl"]
+   delta_means[i]<- mean(weight_boys)- mean(weight_girls)
+ }
> hist(delta_means,las=1,
+       xlab="Différence entre les poids moyens, en grammes (boys - girl)",
+       main="",ylab="Nb. permutations")
> abline(v=real_delta, col="gray",lwd=2,lty=2,)
> p<-sum(delta_means>=real_delta)/1000
```

4. Vous, oui. Mais ce n'est pas la question ici, ce serait trop simple. Le titre aurait du vous donner un indice!



```
> p
[1] 0.064
```

Q16– Au vu de ce graphique et de la p-valeur empirique calculée, que pouvez vous dire sur le poids moyen (au sens de la moyenne harmonique) des bébés garçons par rapport aux bébés filles ? La différence est-elle statistiquement significative, et pourquoi ?

Q17– Le physicien décide de travailler avec une version différente de la moyenne pour comparer garçons et filles. Il choisit la moyenne harmonique, qui se calcule sous , pour un vecteur \mathbf{x} de valeurs numériques, comme `length(x)/sum(1/x)`. Que doit-il modifier dans le code précédent pour réaliser la même approche par permutation avec cette nouvelle statistique ?

Fin de l'examen. Les vacances approchent à grand pas !⁵

5. Si vous êtes arrivés ici et qu'il vous reste du temps, n'oubliez pas que la question 12 était un bonus à traiter en fin d'examen !⁶

6. Si vous avez fini et traité la question 12, je suis impressionné. Et fier de vous. Et j'arrête les notes de bas de page.