TP R sur les tests

Emmanuel Rachelson and Matthieu Vignes

24 janvier 2013, SupAero - ISAE

1 Un petit exercice: la chaleur latente de fusion de la glace

Voici 2 séries de mesures indépendantes de chaleur latente de fusion de la glace (en cal/g) :

Méthode A	79.98 80.04 80.02 80.04 80.03 80.03 80.04 79.97 80.05 80.03 80.02 80.00 80.02
Méthode B	80.02 79.94 79.98 79.97 79.97 80.03 79.95 79.97

1. Lire les données dans R.

2. Comparer les distributions à l'aide de `boxplot()`(s). Conclusions ?

3. Tester pour l'égalité des moyennes avec `t.test()`.Quelles sont les hypothèses de ce test ? Conclusions ?

4. Tester l'hypothèse d'égalité des variances à l'aide de `var.test()`. Conclusions ?

5. Appliquer un *t-test* classique qui suppose l'égalité des variances. Conclusions ?

6. Tous les tests ci-dessus supposent la normalité des deux échantillons. Appliquer un test de rang signé, `wilcox.test()`. Quelles sont les hypothèses de ce test ? Conclusions ?

7. Tester la normalité des données avec `qqplot()` et des tests appropriés (Shapiro-Wilk, Kolmogorov). Conclusions ?

2 Etude d'un scénario complet: étude de la concentration en ozone

2.1 Introduction

La pollution de l'air constitue actuellement une des préoccupations majeures de santé publique. De nombreuses études épidémiologiques ont permis de mettre en évidence l'influence sur la santé de certains composés chimiques comme le dioxyde soufre (SO2), le dioxyde d'azote (NO2), l'ozone (O3) ou des particules en suspension. Des associations de surveillance de la qualité

de l'air (Air Breizh en Bretagne depuis 1994) existent sur tout le territoire métropolitain et mesurent la concentration des polluants. Elles enregistrent également les conditions météorologiques comme la température, la nébulosité, le vent, les chutes de pluie en relation avec les services de Météo France... L'une des missions de ces associations est de construire des modèles de prévision de la concentration en ozone du lendemain à partir des données disponibles du jour : observations et prévisions de Météo France. Plus précisément, il s'agit d'anticiper l'occurrence ou non d'un dépassement légal du pic d'ozone (180 μgr/m3) le lendemain afin d'aider les services préfectoraux à prendre les décisions nécessaires de prévention : confinement des personnes à risque, limitation du trafic routier. Plus modestement, l'objectif de cette étude est de mettre en évidence l'influence de certains paramètres sur la concentration d'ozone (en μgr/m3) et différentes variables observées ou leur prévision. Les 112 données étudiées ont été recueillies à Rennes durant l'été 2001. Elles sont disponibles sur le site du laboratoire de mathématiques appliquées de l'Agrocampus Ouest. Les 13 variables observées sont :

- MaxO3 : Maximum de concentration d'ozone observé sur la journée en μgr/m3
- T9, T12, T15 : Température observée à 9, 12 et 15h
- Ne9, Ne12, Ne15 : Nébulosité observée à 9, 12 et 15h
- Vx9, Vx12, Vx15 : Composante E-O du vent à 9, 12 et 15h
- MaxO3v : Teneur maximum en ozone observée la veille
- vent : orientation du vent à 12h
- pluie : occurrence ou non de précipitations

2.2 Exploration statistique élementaire

- lire les données (fichier `ozone.csv` à récupérer là : `http://carlit.toulouse.inra.fr/wikiz/inde` supprimer la variable inutile `obs`. Aide R: `read.csv2(), summary()`
- description unidimensionnelle : variables qualitatives, quantitiatives. Aide R: `mean, sd, boxplot, hist, barplot, pie`
- description bidimensionnellle : variables quantitatives, variables qualitatives, variables qualitatives vs quantitatives. Aide R: `pairs, plot, table, mosaicplot, boxplot`

2.3 Tests de comparaisons

Important : Lors de l'exécution de chaque test préciser vous bien :

1. la question posée,
2. l'hypothèse (H0) en relation avec la question et associée au test,
3. la p-valeur calculée et la décision du test,
4. la réponse à la question.

2.3.1 Gaussanité

Beaucoup des outils ci-dessous nécessitent de vérifier le caractère gaussien ou non de la distribution. En fait, le nombre important d'observations dans l'échantillon permet de s'affranchir de cette hypothèse mais il est utile de savoir la vérifier et éventuellement de sélectionner la transformation la plus appropriée des données notamment pour les variables de concentration d'ozone.

Normalité d'une distribution : Shapiro-Wilks La droite de Henri ou graphe quantile-quantile donne déjà un aperÃ§u graphique de la normalité de la distribution avant de calculer le test.

```
# qq-plots
qqnorm(ozone$maxO3); qqline(ozone$maxO3,col=2)
qqnorm(log(ozone$maxO3)); qqline(log(ozone$maxO3),col=2)
# Test de shapiro-Wilks
shapiro.test(ozone$maxO3); shapiro.test(log(ozone$maxO3))
```

Le test de Kolmogorov-Smirnov de comparaison à une distribution théorique pourrait également Ã^atre utilisé (`ks.test`).

Les variables transformées sont ajoutées dans la table.

```
ozone=data.frame(ozone,LmaxO3=log(ozone$maxO3), LmaxO3v=log(ozone$maxO3v))
summary(ozone)
```

Intervalle de confiance d'une moyenne : Student Il est important de savoir estimer l'intervalle de confiance d'une moyenne ; celui-ci permet de tester l'égalité de cette moyenne à une valeur théorique selon l'appartenance ou non de cette valeur à l'intervalle. L'effectif étant suffisamment grand, il n'est pas nécessaire de supposer la normalité des données mais la variable transformée la plus "gaussienne" est choisie. L'intervalle de confiance est calculé par défaut avec un seuil à 95% mais ce paramètre peut Ã^atre précisé (`conf.level=.95`) de mÃ^ame que la moyenne théorique testée (`mu=0.0`, par défaut à `0`).

```
t.test(log(ozone$LmaxO3), conf.level=.95)
```

Comparaison de deux variances : Fisher On s'intéresse à l'influence de la présence de pluie sur la concentration en ozone. Tester l'égalité des deux moyennes nécessite de vérifier préalablement plusieurs points :

1. la normalité des distributions dans chaque classe à moins que l'échantillon soit considéré de taille suffisamment grande,
2. le caractère indépendant ou appariés des échantillons,
3. l'égalité ou non des variances à l'intérieur de chaque groupe.

On dispose de deux échantillons indépendants : les jours de pluie et les jours de temps sec. Testons les autres hypothèses.

```
# Normalité des distributions (facultatif)
shapiro.test(ozone[ozone$pluie=="Pluie","LmaxO3"])
shapiro.test(ozone[ozone$pluie=="Sec","LmaxO3"])
# égalité des variances (test de Fisher)
var.test(LmaxO3~pluie,data=ozone)
```

Commenter les résultats.

Comparaison de deux moyennes Le test de comparaison des moyennes à utiliser (Student *vs.* Welsh) dépend du résultat précédent concernant l'égalité des variances.

→ **Echantillons indépendants** Si les variances sont différentes, il s'agit d'un test de Welch.

```
t.test(LmaxO3~pluie,var.equal=F, data=ozone)
```

Dans le cas où elles sont considérées égales, c'est un test de Student.

```
t.test(LmaxO3~pluie,var.equal=T, data=ozone)
```

→ **échantillons appariés** On se propose d'étudier la persistance moyenne de la concentration en comparant la moyenne du jour avec celle de la veille. La mesure étant observée au mÃªme point à deux instants différents, les échantillons sont cette fois appariés.

```
t.test(ozone$maxO3, ozone$maxO3v,paired=TRUE)
```

2.3.2 Cas non-paramétrique

Si l'hypothèse de normalité des distributions n'est pas vérifiée et si l'échantillon est trop réduit, c'est un test non-paramétrique qu'il faut mettre en oeuvre. Les tests non-paramétriques sont basés sur les rangs des observations et donc sur les comparaisons des médianes des échantillons. Une transformation des variables par une fonction monotone (*i.e.* log) qui ne changent pas leur ordonnancement n'a donc pas d'effet sur le calcul d'un test non paramétrique.

Comparaison de deux médianes : Wilcoxon
→ Echantillons indépendants

```
tapply(ozone$LmaxO3, ozone$pluie, median)
wilcox.test(maxO3∼pluie, data=ozone)
```

→ Echantillons appariés

```
median(ozone$LmaxO3 - ozone$LmaxO3v)
wilcox.test(ozone$LmaxO3, ozone$LmaxO3v,paired=TRUE)
```

Comparer avec les résultats des tests paramétriques.

2.4 Tests de liaison: indépendance de 2 variables qualitatives

Le test du χ^2 est adapté à ce problème.

```
chisq.test(table(ozone$pluie,ozone$vent))
```

Rque : un avertissement peut signaler que les effectifs théoriques (sous hypothèse d'indépendance) de certaines cellules sont trop faibles pour justifier des propriétés asymtotiques du test du χ^2. Il est dans ce cas nécessaire de regrouper des modalités.