Les tests statistiques par la pratique !

Emmanuel Rachelson and Matthieu Vignes

23 octobre 2013, SupAero - ISAE

1 Comparaison de la moyenne de 2 échantillons

2 exemples de petits jeux de données:

Longueurs de machoires inférieure (en mm) de 10 chacals mâles et femelles. La variable diffère-t-elle entre les sexes de cette espèce ?

mâles	120	107	110	116	114	111	113	117	114	112
femelles	110	111	107	108	110	105	107	106	111	111

Temps de survie (en h) de souris atteintes de cancer et traitées avec un médicament donné; cette variable dépend elle du type de cancer ?

estomac	124	42	25	45	412	51	1112	46	103	876	146	340	396
poumon	1235	24	1581	1166	40	727	3808	791	1804	3460	719		

Si on suppose les variances égales dans un test de comparaison de moyennes, l'EMV en est $\hat{\sigma^2} = \frac{\sum_{i=1}^{n_1}(x_i - m_1)^2 + \sum_{i=1}^{n_2}(y_i - m_2)^2}{n_1 + n_2 - 2}$. Et la variance de $\bar{X} - \bar{Y}$ est $\sigma^2(1/n_1 + 1/n_2)$. La variable normalisée définie par :

$$t = \frac{m_1 - m_2}{\hat{\sigma^2}(1/n_1 + 1/n_2)}$$

suit une loi de Student à $n_1 + n_2 - 2$ ddl.

```
x <- seq(-3, 3, le = 100)
par(mfrow=c(1,1))
plot(x, dnorm(x,-0.5), type = "l", ylim = c(-0.3,0.4))
lines(x, dnorm(x,0.5), type = "l")
y1 <- rnorm(12,-0.5)
y2 <- rnorm(10,0.5)
abline(h = c(0,-0.1,-0.2))
points (y1, rep(-0.1,12))
points(y2, rep(-0.2,10))
```

La loi de la différence des moyennes est :

```
plot(x, dt(x,20), type = "l"); plot(x, dnorm(x), type = "o")
```

Si l'alternative est $\mu_1 > \mu_2$, on attend des valeurs positives pour $m_1 - m_2$ et on rejette (H0) avec le risque $P(T > t)$. Si l'alternative est $\mu_1 < \mu_2$, des

valeurs négatives pour $m_1 - m_2$ sont attendues et on rejette (H0) avec un risque $P(T < t)$. Enfin si (H1) $\mu_1 \neq \mu_2$, on attend de fortes valeurs positives ou négatives pour $m_1 - m_2$ et on rejette (H0) avec $2\,P(T < t)$.

Si on applique la procédure à la situation 1, on obtient:

```
x1 <- c(120,107,110,116,114,111,113,117,114,112)
x2 <- c(110,111,107,108,110,105,107,106,111,111)
m1 <- mean(x1); m2 <- mean(x2)
v <- (9*var(x1) + 9*var(x2))/18 ; t <- (m1-m2)/sqrt(v*(1/10 + 1/10))
tt0 <- t.test(x1, x2, var.eq = T)
tt0
```

Note: la proba pour avoir une différence sup à 3.48 ou inf. à -3.48 est `1-pt(3.48,18)` ici; (H0) est rejetée au risque de première espèce de 3 pr 1000.

Appliquée à la situation 2, cette procédure donne :

```
y1 <- c(124, 42, 25, 45, 412, 51, 1112, 46, 103, 876, 146, 340,
396)
y2 <- c(1235, 24, 1581, 1166, 40, 727, 3808, 791, 1804, 3460, 719)
tt0 <- t.test(y1, y2, var.eq = T)
tt0
```

Le test de Wilcoxon (Mann-Whitney)

On pourrait croire que les 2 situations sont identiques. En fait non :

```
hist(c(x1,x2), nclass = 8, col=grey(0.7), main="longueur de la machoire")
hist(c(y1,y2), nclass = 8, col=grey(0.7), main="temps de survie")
```

Des commentaires sur la syméétrie des distributions ? Et l'hypothèse de normalité alors ? Donc rejeter l'hypothèse de normalité des moyennes n'a pas de sens. On va donc adopter une stratégie *libre de distribution*. Le plus simple est le test de Wilcoxon (aussi appelé Mann-Whitney). On va réunir les 2 échantillons:

```
ytot <- c(y1,y2)
```

Puis calculer les rangs des valeurs des variables et on sépare les rangs des 2 groupes :

```
rtot <- rank(ytot)
n1 <- length(y1); n2 <- length(y2)
r1 <- rtot[1:n1]; r2 <- rtot[(n1+1):n2]
```

Si les individus des 2 groupes proviennent de la même population, les rangs d'un groupe sont tirés au hasard parmi les 24 premiers entiers. Si les moyennes des 2 échantillons ne sont pas égales, les rangs du premier groupe seront soit trop grands soit trop petits. On utilise comme statistique de test la somme des rangs SR. On peut raisonner de manière symétrique sur les 2 groupes; pour fixer les idées, on va travailler sur le premier, d'effectif m et soit n l'effectif total. Dans $\binom{n}{m}$, si $m \geq 10$ et $n - m \geq 10$, SR suit approx-

imativement une loi normale de moyenne $\frac{m(n+1)}{2}$ et variance $\frac{m(n-m)(n+1)}{12}$.
Il y a des soucis en cas d'*ex aequo* mais des correctifs existent. Les logiciels étudient souvent $SR - m(m+1)/2$.

```
sr <- sum(r1) # 122 est la somme des rangs de l'echantillon 1
esr <- (n1 * (n1 + n2 + 1))/2; vsr <- (n1 * n2 * (n1 + n2 + 1))/12
t0 <- (sr - esr)/sqrt(vsr)
2 * (pnorm(t0))
u <- sr - (n1 * (n1+1))/2 # 31 :  valeur de la stat.  de Mann-Whitney
...
wilcox.test(y1, y2, exact = F, correct = F)
wilcox.test(y1, y2, exact = T) # ou mieux la valeur exacte
```

L'approximation est bien justifiée. Sur l'exemple de longueur des mâchoires de chacal... On refait et ?

A retenir : le test non paramétrique est utilisable dans tous les cas et donne des résultats plus solides.

2 Comparaison de s échantillons

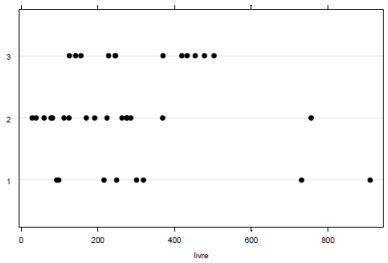
2.1 Test de Kruskal-Wallis

Dans la bibliothèque de Peter Sprent 6, il y a des livres de voyage, des ouvrages généraux et des livres de statistiques . On a trois échantillons. Le premier est celui des livres de voyage. Ils ont respectivement 93, 98, 216, 249, 301, 319, 731 et 910 pages. Le second est celui des livres généraux. Ils ont 29, 39, 60, 78, 82, 112, 125, 170, 192, 224, 263, 275, 276, 286, 369 et 756 pages. Le troisième est celui des livres de statistiques. Ils ont 126, 142, 156, 228, 245, 246, 370, 419, 433, 454, 478 et 503 pages. La question est "Ces échantillons proviennent-ils d'une même population ou au contraire un au moins des trois échantillons présentent une originalité en ce qui concerne la taille moyenne ?".

Je suis sympa ;) :

```
 livre <- c(93,98,216,249,301,319,731,910,29,39,60,78,82,112,125,
+170,192,224,263,275,276,286,369,756,126,142,156,228,245,246,370,
+419,433,454,478,503)
```

Et créer un vecteur `groupe` (penser à `rep()`). Créer une belle représentation comme :



Si SRj est la somme des rangs de l'échantillon j, la variable T définie par

$$T := \frac{12 \sum_{j=1}^{s} SRj^2/n_j}{n(n+1)} - 3(n+1)$$

suit une loi du χ^2 à $s-1$ ddl. Cela suffit pour éxécuter le test de Kruskal-Wallis et conclure :

```
kruskal.test(livre,groupe)
```

***A parte* sur le χ^2** La loi du χ^2 à m ddl est définie comme la somme de m carrés de loi normale standard indépendantes. Les densités typiques sont de la forme :

```
x0 = seq(0,30,le=100)
y1 = dchisq(x0,3)
y2 = dchisq(x0,5)
y3 = dchisq(x0,10)
y4 = dchisq(x0,15)
y5 = dchisq(x0,20)
plot(x0, y1, type="n", xlab="x", ylab="Chi2 density")
lines(x0, y1, lty = 1)
lines(x0, y2, lty = 2)
lines(x0, y3, lty = 3)
lines(x0, y4, lty = 4)
lines(x0, y5, lty = 5)
l0 = c("Khi2 3 ddll","Khi2 5 ddll","Khi2 10 ddll","Khi2 15 ddll",
+"Khi2 20 ddll"); legend(10,0.2,l0, lty=1:5)
```

2.2 Comparer les variances

Dans trois groupes de TD, les notes de contrôle continu sont :

```
x1 <- c(14.9,12.0,9.5,7.3,8.4,9.8,11.0,13.8,14.3,5.0,4.4,14.3,13.7,18.0,12.4)
x2 <- c(10.9,10.1,10.0,12.2,10.0,11.1,10.3,9.5,9.6,10.0,10.9,11.2)
x3 <- c(13.0,12.1,8.7,10.9,12.7,9.5,10.5,12.2,16.0,10.3,9.6,10.9,7.3,9.8)
```

Les enseignants se réunissent pour vérifier qu'il n'y a pas de différence notable entre leurs notations.

```
kruskal.test(c(x1, x2, x3), gtd)
```

Tout va bien ? "PAS DU TOUT" s'insurge le représentant des étudiants. Faites donc le bon dessin :

```
dotplot(gtd~c(x1,x2,x3), cex=1.5)
```

Discussion stérile: "Vous voyez bien que les amplitudes de notation diffèrent grandement entre les groupes ?!". "Mais non, c'est le hasard...". "Impossible !". "Et bien prouvez le !"

```
median(x)
kruskal.test(abs(x - 10.9), gtd)
dotplot(gtd~abs(x - 10.9), cex = 1.5)
```

Moralité : attention avant de traverser, un test peut en cacher un autre...

3 ANOVA

3.1 Analyse de variance à un facteur

Trois machines sont réglées pour produire des pièces identiques dont la caractéristique X est de loi $\mathcal{N}(\mu, \sigma^2)$. POur s'assurer qu'elles ne sont pas déréglées (elles le sont si les moyennes sont distinctes), on prélève un échantillon produit par chaque machine. Le tableau des valeurs obtenues est le suivant:

i	1	2	3	4	5
x_1	5	7	6	9	13
x_2	8	14	7	12	9
x_3	14	15	17	18	11

Alors ?

3.2 Analyse de variance à deux facteurs

Des ampoules sont fabriquées en utilisant 4 types de filaments (facteur A à 4 modalités) et 4 types de gaz (facteur B à 4 modalités aussi). On souhaite évaluer l'effet de chacun des facteurs sur la durée de vie des ampoules. On supposera qu'il n'y a pas d'interaction...Le tableau des durées de vie est le suivant :

	B1	B2	B3	B4
A1	44	22	36	34
A2	47	43	41	53
A3	0	9	10	17
A4	36	14	1	34