Probabilités et statistique ELEC, MICRO, MX
Été 2006 Dr. Diego Kuonen

Exercices du 8 juin 2006 **TP 2**

La section MX se limite aux exercices suivis d'une étoile ***.***

Pour commencer, copier le fichier `TP2.r` qui se trouve sur la page du site web et charger les jeux de données avec

```
> source('c:\temp\TP2.r')
```

ou bien taper directement

```
> source('http://statwww.epfl.ch/teaching/data/TP2.r')
```

Exercice 1. Test de Student *
On a contrôlé 10 compteurs d'électricité nouvellement fabriqués:

983 1002 998 996 1002 983 994 991 1005 986

On aimerait savoir s'il y a un écart systématique entre la valeur standard 1000 et les compteurs fabriqués. En d'autres termes, on va tester si la moyenne, disons μ, des données égale ou diffère de 1000. Pour cela on va utiliser un modèle normal, cela signifie que l'on suppose que nos données sont les réalisations de variables aléatoires $X_1, \ldots, X_{10}$ indépendantes et identiquement distribuées selon une loi $\mathcal{N}(\mu, \sigma^2)$.

a) On doit tout d'abord contrôler la validité de cette hypothèse. Pour cela on trace un QQ-plot des observations contre la loi normale.

```
> qqnorm(compteurs)
> abline(mean(compteurs), sd(compteurs))
```

Que conclure sur la base de ce graphique?

b) On effectue le test

$$\mathcal{H}_0 : \mu = 1000 \quad \text{contre} \quad \mathcal{H}_1 : \mu \neq 1000.$$

Il a été vu au cours que le test à utiliser est celui de Student (appelé aussi test t) basé sur la statistique de test

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

qui, sous l'hypothèse $\mathcal{H}_0$, suit une loi de Student t_{n-1}.

i) Effectuer ce test à un niveau $\alpha = 5\%$ et conclure.

```
> t.test(compteurs, mu=1000, alternative='two.sided', conf.level=0.95)
```

ii) Effectuer ce test à un niveau $\alpha = 10\%$ et conclure.

```
> t.test(compteurs, mu=1000, alternative='two.sided', conf.level=0.90)
```

iii) Effectuer ce test à un niveau $\alpha = 1\%$ et conclure.

```
> t.test(compteurs, mu=1000, alternative='two.sided', conf.level=0.99)
```

c) Quelle est l'influence du choix du niveau du test? Discuter la dualité entre l'intervalle de confiance et le test.

Exercice 2. Têtes de vis
Le tableau suivant représente la distribution des diamètres de têtes de vis fabriquées par un atelier suisse.

Effectif	2	6	8	15	42	68	49	25	18	12	4	1
Diamètre (cm)	0.725	0.726	0.727	0.728	0.729	0.730	0.731	0.732	0.733	0.734	0.735	0.736

Selon les nouvelles normes européennes, la moyenne des têtes de vis doit être 0.73 cm. Les mesures obtenues fournissent-elles suffisamment d'évidence, au niveau de $\alpha = 1\%$, pour rejeter l'hypothèse que les têtes ont la moyenne souhaitée?

```
> t.test(tete2vis, mu=0.73, alternative='two.sided', conf.level=0.99)
```

Exercice 3. Test de Student apparié
Douze paires d'agneaux jumeaux sont sélectionnées pour comparer deux régimes I et II. Dans chaque paire les agneaux ont été nourris avec des régimes différents. Les poids des agneaux après huit mois, sont les suivants:

paire	1	2	3	4	5	6	7	8	9	10	11	12
régime I	91	97	102	93	95	90	99	119	86	97	87	97
régime II	90	99	106	95	97	93	101	116	89	100	89	95

On souhaite tester $\mathcal{H}_0$: "les régimes I et II sont équivalents" contre $\mathcal{H}_1$: "le régime II est plus riche que le régime I" au niveau de $\alpha = 5\%$.
Appliquez le test de Student, en tenant compte que les paires d'agneaux sont des jumeaux (données appariées).

```
> t.test(x=agneau[,1], y=agneau[,2], alternative='less', mu=0, paired=TRUE, conf.level=0.95)
```

Exercice 4. Une analyse de données *
Les données `iqnd` et `iqd` sont issues d'une expérience visant à mettre en évidence l'effet éventuel d'un état dépressif d'une mère sur le score d'un test de QI de son enfant. Dans cet exercice, on va analyser ce jeu de données.

a) Discuter la conclusion de chaque résultat des lignes de code suivantes.

```
> boxplot(iqnd, iqd, names=c('Meres non-depressives','Meres depressives'),
  ylab='Scores de QI')
> par(mfrow=c(2,1))
> hist(iqnd, xlab='QI des enfants de meres non-depressives', ylab='Frequence')
> hist(iqd, xlab='QI des enfants de meres depressives', ylab='Frequence')
> qqnorm(iqnd, ylab='Scores de QI')
> qqline(iqnd)
> title('QQ-plot normal des QI des enfants de meres non-depressives')
> qqnorm(iqd, ylab='Scores de QI')
> qqline(iqd)
> title('QQ-plot normal des QI des enfants de meres depressives')
```

b) Maintenant effectuer un test de Student de niveau 5% sur la différence des moyennes des QI des deux types d'enfants.

```
> t.test(iqnd, iqd, alternative='two.sided', mu=0, paired=FALSE, conf.level=0.95)
```

Conclure.

c) Les variances des échantillons sont très différentes,

```
> var(iqnd)
> var(iqd)
```

et il se trouve que cela peut être dû aux valeurs aberrantes révélées par les boxplots effectués au point **a)**. Pour vérifier leur effet, on les retire des jeux de données et on calcule à nouveau la variance.

```
> iqnd
> iqnd = iqnd[iqnd > 50]
> iqnd
> iqd
> iqd = iqd[iqd > 50]
> iqd
> var(iqnd)
> var(iqd)
```

Les variances sont maintenant beaucoup plus proches, on va tester leur égalité en utilisant un test F (dernier test du tableau qui a été distribué).

```
> var.test(iqnd, iqd, ratio=1, alternative='two.sided', conf.level=0.95)
```

d) On effectue à nouveau un test t, sans les valeurs aberrantes. En utilisant le résultat précédent, décider quelle ligne de code est la plus appropriée.

```
> t.test(iqnd, iqd, alternative='two.sided', mu=0, paired=FALSE, conf.level=0.95)
> t.test(iqnd, iqd, alternative='two.sided', mu=0, paired=FALSE, var.equal=TRUE,
  conf.level=0.95)
```

e) *Selon vous*, est-il justifié de retirer les valeurs aberrantes des jeux de données?

Exercice 5. Régression linéaire *

Une analyse de marche sur 30 m, effectuée sur $n = 16$ personnes âgées, nous permet de mesurer la longueur moyenne des pas (x_i) et la vitesse moyenne de marche (y_i), $i = 1, \ldots, n$. Les données sont montrées dans le tableau suivant:

x_i (m)	0.92	0.92	0.60	0.50	0.99	0.79	0.75	0.97
y_i (m/s)	0.56	0.56	0.42	0.39	0.42	0.40	0.32	0.69

x_i (m)	0.48	1.15	0.90	0.99	0.85	0.77	0.90	0.68
y_i (m/s)	0.23	1.03	0.78	0.68	0.60	0.68	0.64	0.58

On cherche à mettre en évidence la relation qui lie y à x.

a) On commence par faire une inspection graphique des données avec un scatterplot.

```
> plot(marche$x, marche$y, xlab='Longueur des pas', ylab='Vitesse moyenne de marche')
```

b) On peut également calculer la corrélation pour mesurer la dépendance linéaire entre x et y.

```
> cor(marche$x, marche$y)
```

c) La commande suivante permet d'ajuster la "meilleure" droite

$$y_i \approx a + bx_i, \quad i = 1, \ldots, 16.$$

```
> marche.mod = lm(marche$y ~ 1 + marche$x, data=marche)
> summary(marche.mod)
> plot(marche$x, marche$y, xlab='Longueur des pas', ylab='Vitesse moyenne de marche')
> abline(marche.mod, col=2)
```

On peut également ajuster la "meilleure" parabole

$$y_i \approx a + bx_i + cx_i^2, \quad i = 1, \ldots, 16.$$

```
> marche.mod2 = lm(marche$y ~ 1 + marche$x + I(marche$x^2), data=marche)
> summary(marche.mod2)
> plot(marche$x, marche$y, xlab='Longueur des pas', ylab='Vitesse moyenne de marche')
> absc = seq(from=par('usr')[1], to=par('usr')[2], length=1000)
> parabole = marche.mod2$coef[1] + absc*marche.mod2$coef[2] + absc^2*marche.mod2$coef[3]
> points(absc, parabole, type='l', col=2)
```