

Modèle linéaire gaussien

Didier CHAUVEAU et Jean-François DELMAS

October 10, 2017

Contents

1	Peut-on augmenter la quantité de pluie?	1
1.1	La problématique	1
1.2	Les données	2
1.3	Le modèle	2
1.4	Comparaison des moyennes	3
1.5	(FACULTATIF) COMPARAISON DES VARIANCES DES 2 ÉCHANTILLONS GAUSSIENS	5
2	(FACULTATIF) ANALYSE DE LA VARIANCE À UN FACTEUR	6
2.1	Les données	6
2.2	Le modèle	7
3	Régression linéaire	9
3.1	Les données	9
3.2	Régression linéaire simple	10
3.3	Régression linéaire multiple	11

1 Peut-on augmenter la quantité de pluie?

1.1 La problématique

Il existe deux types de nuages qui donnent lieu à des précipitations: les nuages chauds et les nuages froids. Ces derniers possèdent une température maximale de l'ordre de -10°C à -25°C . Ils sont composés de cristaux de glace et de gouttelettes d'eau. Ces gouttelettes d'eau subsistent alors que la température ambiante est inférieure à la température de fusion. On parle d'eau surfondue. Leur état est instable. De fait, quand une particule de glace rencontre une gouttelette d'eau, elles s'aggrègent pour ne former qu'une seule particule de glace. Les particules de glace, plus lourdes que les gouttelettes, tombent sous l'action de la gravité. Enfin si les températures des couches d'air inférieures sont suffisamment élevées, les particules de glace fondent au cours de leur chute formant ainsi de la pluie.

En l'absence d'un nombre suffisant de cristaux de glace pour initier le phénomène décrit ci-dessus, on peut injecter dans le nuage froid des particules qui ont une structure cristalline proche de la glace, par exemple de l'iodure d'argent (environ 100 à 1000 grammes par nuage). Autour de ces particules, on observe la formation de cristaux de glace, ce qui permet, on l'espère, de déclencher ou d'augmenter les précipitations. Il s'agit de l'ensemencement des nuages. Signalons que cette méthode est également utilisée pour limiter le risque de grêle.

Il est évident que la possibilité de modifier ainsi les précipitations présente un grand intérêt pour l'agriculture. De nombreuses études ont été et sont encore consacrées à l'étude de l'efficacité de l'ensemencement des nuages dans divers pays. L'étude de cette efficacité est cruciale et délicate. Le débat est encore d'actualité.

1.2 Les données

On dispose des données concernant l'ensemencement par iodure d'argent des nuages en Floride (1975)¹.

Les volumes de pluie déversés en 10^7m^3 (cf. les deux tableaux ci-dessous) concernent 23 jours similaires dont $n_1 = 11$ jours avec ensemencement correspondant aux réalisations des variables aléatoires X_1, \dots, X_{n_1} et $n_2 = 12$ jours sans ensemencement correspondant aux réalisations des variables aléatoires Y_1, \dots, Y_{n_2} .

i	1	2	3	4	5	6	7	8	9	10	11
X_i	7.45	4.70	7.30	4.05	4.46	9.70	15.10	8.51	8.13	2.20	2.16

Table 1: Volume de pluie en 10^7m^3 déversée avec ensemencement

j	1	2	3	4	5	6	7	8	9	10	11	12
Y_j	15.53	10.39	4.50	3.44	5.70	8.24	6.73	6.21	7.58	4.17	1.09	3.50

Table 2: Volume de pluie en 10^7m^3 déversée sans ensemencement

$X=[7.45, 4.70, 7.30, 4.05, 4.46, 9.70, 15.10, 8.51, 8.13, 2.20, 2.16]$;
 $Y=[15.53, 10.39, 4.50, 3.44, 5.70, 8.24, 6.73, 6.21, 7.58, 4.17, 1.09, 3.50]$;

1.3 Le modèle

On suppose que les volumes de pluies suivent une loi gaussienne et que l'effet de l'ensemencement ne modifie pas la variance des lois gaussiennes.

¹William L. Woodley, Joanne Simpson, Ronald Biondini, Joyce Berkeley, *Rainfall results, 1970-1975: Florida Area Cumulus Experiment*, Science, Vol 195, pp. 735-742, February 1977.

Ceci revient à dire dans ce contexte paramétrique que les observations sont constituées de deux échantillons gaussiens indépendants

$$\begin{aligned} X_1, \dots, X_{n_1} & \text{ i.i.d. } \mathcal{N}(\mu_1, \sigma^2), \\ Y_1, \dots, Y_{n_2} & \text{ i.i.d. } \mathcal{N}(\mu_2, \sigma^2). \end{aligned}$$

On souhaite tester l'hypothèse nulle H_0 : “le procédé n'a pas d'effet” contre “le procédé produit une augmentation significative de la quantité de pluie”, autrement dit

$$H_0 = \{\mu_1 = \mu_2\} \quad \text{contre} \quad H_1 = \{\mu_1 > \mu_2\}.$$

Il est indispensable de s'assurer d'abord que les hypothèses de modèle que l'on a faites sont raisonnables:

1. Les données peuvent-elles être considérées comme des réalisations de lois gaussienne pour les paramètres (moyenne, variance) appropriés?
2. Peut-on considérer que les variances sont les mêmes?

Des éléments de réponses au point (1) seront abordés ultérieurement dans le cours. Le point (2) est abordé dans le paragraphe facultatif 1.5.

Nous admettons donc dans un premier temps ces deux hypothèses.

1.4 Comparaison des moyennes

Afin de se faire une idée des données et de l'écart entre les deux populations, on peut commencer par calculer les moyennes empiriques

$$\bar{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i, \quad \bar{Y} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i,$$

ainsi que les variances empiriques

$$\begin{aligned} S_X^2 &= \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2 = \frac{n_1}{n_1 - 1} \left[\frac{1}{n_1} \sum_{i=1}^{n_1} X_i^2 - (\bar{X})^2 \right], \\ S_Y^2 &= \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 = \frac{n_2}{n_2 - 1} \left[\frac{1}{n_2} \sum_{i=1}^{n_2} Y_i^2 - (\bar{Y})^2 \right]. \end{aligned}$$

```
// taille des echantillons
n1 = length(X);
n2 = length(Y);
// calcul des moyennes
Xbar=sum(X)/n1
Ybar=sum(Y)/n2
```

```
// calcul de la somme des carres
SSX=sum(X.^2);
SSY=sum(Y.^2);
// calcul des variances empiriques sans biais
SX2 = SSX/(n1-1)- n1/(n1-1)*Xbar^2
SY2 = SSY/(n2-1)- n2/(n2-1)*Ybar^2
```

On rappelle que les variables aléatoires \bar{X} (resp. \bar{Y}) et $(n_1-1)S_X^2/\sigma^2$ (resp. $(n_2-1)S_Y^2/\sigma^2$) sont indépendantes et de loi $\mathcal{N}(\mu_1, \sigma^2/n_1)$ (resp. $\mathcal{N}(\mu_2, \sigma^2/n_2)$) et $\chi^2(n_1-1)$ (resp. $\chi^2(n_2-1)$). Comme S_X^2 et S_Y^2 sont indépendants, on en déduit que la loi de $\frac{(n_1-1)S_X^2}{\sigma^2} + \frac{(n_2-1)S_Y^2}{\sigma^2}$ est la loi du χ^2 à $n_1 + n_2 - 2$ degrés de liberté.

Sous H_0 , la loi de $\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{(\bar{X} - \bar{Y})}{\sigma}$ est la loi gaussienne centrée réduite. On en déduit que sous H_0 ,

$$T = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{(n_1-1)S_X^2 + (n_2-1)S_Y^2}{n_1 + n_2 - 2}}}$$

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

Comme sous H_1 , $\bar{X} - \bar{Y}$ converge p.s. vers $\mu_1 - \mu_2 > 0$ quand $\min(n_1, n_2) \rightarrow \infty$, et que $\lim_{\min(n_1, n_2) \rightarrow \infty} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = +\infty$, on en déduit que sous H_1 , la statistique T diverge vers $+\infty$ quand $\min(n_1, n_2) \rightarrow \infty$.

Question 1 Construire, à partir de la statistique T , un test d'égalité des moyennes en déterminant sa région critique, la valeur critique au niveau α et la p -valeur.

Question 2 En utilisant la fonction `ttest2` définie ci-dessous, calculer numériquement la valeur de la statistique de test, la valeur critique au niveau 5%, et la p -valeur pour les données de pluie. Conclusion.

La fonction `ttest2` (`ttest2(n1, Xbar, SX2, n2, Ybar, SY2, alpha)`) retourne la valeur de la statistique T observée, la valeur critique associée au niveau α et la p -valeur du test suggéré dans la question précédente. Cliquer sur le lien ci-dessus pour obtenir le code de la fonction, le sauvegarder, sous le nom `ttest2.sce`, dans le répertoire où vous utilisez scilab. Pour charger la fonction, utiliser la commande: `getf 'ttest2.sce'`. Pour appliquer la fonction, utiliser la commande:

```
ttest2(n1, Xbar, SX2, n2, Ybar, SY2, alpha)
```

On donne quelques indications pour la compréhension du code Scilab de `ttest2` qui utilise la fonction de répartition ("cumulative distribution function" en anglais) de la loi de Student:

- $c = \text{cdf}t("T", k, p, 1-p)$ donne à c la valeur du quantile d'ordre p de la loi de Student de paramètre k : $p = \mathbb{P}(T \leq c)$, où T suit la loi de Student de paramètre k .
- $p = \text{cdf}t("PQ", t, k)$ donne à p la valeur $p = \mathbb{P}(T \leq t)$, où T suit la loi de Student de paramètre k .

Pour avoir plus d'information sur la fonction `cdf`t, on peut consulter le manuel en utilisant la commande `help cdf`t.

Le but de ce qui suit est d'observer les résultats du tests quand les valeurs observées ou la taille de l'échantillon varient.

Question 3

1. Faire varier \bar{Y} , les autres valeurs étant fixées. Pour quelles valeurs de \bar{Y} rejetez vous H_0 ?
2. Avec la valeur de \bar{Y} ainsi déterminée, qu'observez vous si S_Y^2 augmente, si S_Y^2 diminue?
3. En reprenant les valeurs numériques du problème, qu'observez vous si n_1 et n_2 augmentent, si n_1 et n_2 diminuent jusqu'aux valeurs limites?

Question 4 Reprendre les questions 1 et 2, si l'hypothèse nulle est $\{\mu_1 \leq \mu_2\}$.

1.5 (FACULTATIF) COMPARAISON DES VARIANCES DES 2 ÉCHANTILLONS GAUSSIENS

Pour s'assurer du point (2), on peut tout d'abord supposer que ces variances sont différentes, i.e.

$$\begin{aligned} X_i &\sim \mathcal{N}(\mu_1, \sigma_1^2), & i = 1, \dots, n_1, \\ Y_j &\sim \mathcal{N}(\mu_2, \sigma_2^2), & j = 1, \dots, n_2, \end{aligned}$$

puis construire un test d'égalité des variances, afin de voir si la différence sur les variances (ou les écart-types) observée est significative.

Question 5 (Facultatif) Quelle est la loi de la statistique $F = S_X^2/S_Y^2$ sous l'hypothèse nulle d'égalité des variances? Quel est son comportement sous H_1 , quand $n_1 \rightarrow \infty$ et $n_2 \rightarrow \infty$?

Question 6 (Facultatif) Construire, à partir de la statistique F , un test d'égalité des variances. Déterminer la région critique du test, évaluer la valeur critique au niveau $\alpha = 5\%$ et la p -valeur pour les données de pluie? Conclusion.

On donne les commandes Scilab suivantes concernant la fonction de répartition de la loi de Fisher:

- $c = \text{cdf}(\text{"F"}, k1, k2, p, 1-p)$ donne à c la valeur du quantile d'ordre p de la loi de Fisher de paramètres $(k1, k2)$: $p = \mathbb{P}(F \leq c)$, où F suit la loi de Fisher de paramètres $(k1, k2)$.
- $[p, q] = \text{cdf}(\text{"PQ"}, f, k1, k2)$ donne à p et q les valeurs $p = 1 - q = \mathbb{P}(F \leq f)$, où F suit la loi de Fisher de paramètres $(k1, k2)$.

Pour avoir plus d'information sur la fonction `cdf`, on peut consulter le manuel en utilisant la commande `help cdf`.

2 (FACULTATIF) ANALYSE DE LA VARIANCE À UN FACTEUR

2.1 Les données

On dispose de données (réelles) relatives à 6 jeux de notes pour un même contrôle, corrigé par un unique correcteur, mais concernant six petites classes (et donc six enseignants) différentes.

On désire savoir si les résultats des petites classes sont significativement différents.

Note	Classe	Note	Classe	Note	Classe
15.50	1	15.50	1	12.00	5
16.00	3	11.75	2	18.50	2
14.00	5	18.50	4	8.00	6
13.75	2	8.00	5	10.50	3
17.00	2	15.00	4	7.00	4
11.00	6	12.50	2	11.00	6
16.50	1	15.50	4	12.50	5
12.75	5	19.00	1	15.00	5
15.00	6	19.00	1	12.50	4
13.50	5	9.50	2	16.50	2
14.00	2	8.00	5	16.50	1
19.50	1	16.50	1	18.00	4
15.50	5	9.50	6	9.50	5
14.50	6	19.00	1	13.50	6
13.00	2	17.00	6	13.00	3
18.00	4	17.00	5	16.00	2
15.00	5	9.50	5	16.00	4
14.50	6	15.50	2	12.00	3
14.50	5	15.00	4	18.50	4
18.50	3	17.00	2	12.00	3
17.50	1	19.50	3	13.00	6
18.00	3	16.50	6	14.50	3
11.00	3	7.00	4	12.00	3
10.50	5	8.50	2	19.50	1
10.00	1	12.50	4	12.00	4
16.00	3	18.75	6	10.50	1
12.25	4	16.00	4	15.00	6
9.50	1	19.50	3	15.50	2
12.00	5	17.50	1	14.25	3
13.50	3	14.50	4	14.00	2
19.50	1	9.00	6	15.50	3
18.50	6				

2.2 Le modèle

On fait l'hypothèse que les notes sont distribuées suivant une loi gaussienne de même variance, et de moyenne dépendant de l'enseignant. Le modèle s'écrit donc

$$X_{i,j} = m_i + \varepsilon_{i,j}, \quad i = 1, \dots, 6, \quad j = 1, \dots, n_i,$$

où m_i est la note moyenne (inconnue) liée à l'enseignant i , et les $\varepsilon_{i,j}$ sont des v.a. indépendantes de loi $\mathcal{N}(0, \sigma^2)$, où $\sigma^2 > 0$ est inconnu également.

```
// notes
X=[
15.50; 15.50; 12.00; 16.00; 11.75; 18.50; 14.00; 18.50; 8.00; 13.75;
8.00; 10.50; 17.00; 15.00; 7.00; 11.00; 12.50; 11.00; 16.50; 15.50;
12.50; 12.75; 19.00; 15.00; 15.00; 19.00; 12.50; 13.50; 9.50; 16.50;
14.00; 8.00; 16.50; 19.50; 16.50; 18.00; 15.50; 9.50; 9.50; 14.50;
19.00; 13.50; 13.00; 17.00; 13.00; 18.00; 17.00; 16.00; 15.00; 9.50;
16.00; 14.50; 15.50; 12.00; 14.50; 15.00; 18.50; 18.50; 17.00; 12.00;
17.50; 19.50; 13.00; 18.00; 16.50; 14.50; 11.00; 7.00; 12.00; 10.50;
8.50; 19.50; 10.00; 12.50; 12.00; 16.00; 18.75; 10.50; 12.25; 16.00;
15.00; 9.50; 19.50; 15.50; 12.00; 17.50; 14.25; 13.50; 14.50; 14.00;
19.50; 9.00; 15.50; 18.50];
// groupe
Grp=[
1; 1; 5; 3; 2; 2; 5; 4; 6; 2; 5; 3; 2; 4; 4; 6; 2; 6; 1; 4; 5; 5; 1; 5;
6; 1; 4; 5; 2; 2; 2; 5; 1; 1; 1; 4; 5; 6; 5; 6; 1; 6; 2; 6; 3; 4; 5; 2;
5; 5; 4; 6; 2; 3; 5; 4; 4; 3; 2; 3; 1; 3; 6; 3; 6; 3; 3; 4; 3; 5; 2; 1;
1; 4; 4; 3; 6; 1; 4; 4; 6; 1; 3; 2; 5; 1; 3; 3; 4; 2; 1; 6; 3; 6];
```

On peut résumer les données.

```
// nombre de groupes
k = max(Grp);
// calcul de l'effectif n(i), de la moyenne moy(i)
// et de la variance s2(i) du groupe i
for i=1:k,
n(i) = length(X(Grp==i));
moy(i) = sum(X(Grp==i))/n(i);
ss(i)=sum(X(Grp==i).^2);
s2(i)=ss(i)/(n(i)-1)- n(i)/(n(i)-1)*moy(i)^2;
end;
[n,moy,s2] // effectif, moyenne et variance par groupe
```

Question 7 Quelles sont les estimations des m_i , $i = 1, \dots, 6$? Quelle est l'estimation de σ^2 ?

Question 8 Peut-on considérer qu'il existe une différence significative entre les notes des classes, dues aux qualités pédagogiques des enseignants? Effectuer le test de l'hypothèse nulle H_0 : "les enseignants ne présentent pas de différence" contre "c'est faux". Écrire la table d'analyse de la variance et conclure.

On pourra utiliser les commandes suivantes:

```
// calcul de la moyenne generale
mg=sum(X)/length(X);
// calcul des sommes de carres
SSM = sum(n.*(moy-mg).^2);
SSE = sum(s2.*(n-1));
```

La fonction `anova(n,moy,s2,alpha)` retourne la table d'analyse de la variance associée. Les paramètres sont `n`, vecteur des tailles de chacun des groupes; `moy`, vecteur des moyennes de chacun des groupes; `s2`, vecteur des variances de chacun des groupes; `alpha`, niveau du test d'analyse de la variance. Cliquer sur le lien ci-dessus pour obtenir le code de la fonction, le sauvegarder sous le nom `anova.sce` dans le répertoire où vous utilisez scilab. Pour charger la fonction, utiliser la commande: `getf 'anova.sce'`. Pour appliquer la fonction, utiliser la commande:

```
anova(n,moy,s2, alpha)
```

Le but de ce qui suit est d'observer les résultats du tests quand les valeurs observées ou la taille de l'échantillon varient.

Question 9

1. *Faire varier le vecteur `moy`, les autres valeurs étant fixées. Pour quelles valeurs de `moy` rejetez vous H_0 ?*
2. *Qu'observez vous si le vecteur, `s2`, des variances augmente, s'il diminue?*
3. *En reprenant les valeurs numériques du problème, qu'observez vous si le vecteur, `n`, des tailles des groupes augmente ou diminue?*

Question 10 *Si l'on rejette H_0 , il est naturel de déterminer les classes qui sont significativement différentes.*

1. *Proposer des intervalles de confiance pour les notes moyennes pour chaque classe.*
2. *Proposer un test de l'hypothèse nulle*

$$H_0 = \{m_i = m_j\} \quad \text{contre} \quad H_1 = \{m_i \neq m_j\},$$

fondé sur la loi de Student, et l'utiliser afin de comparer quelques couples de moyennes.

3 Régression linéaire

3.1 Les données

L'exemple utilisé a déjà été proposé dans les exercices sur le chapitre 2. On dispose de données concernant l'âge (X_1), le kilométrage en milliers de kms (X_2), et le prix en milliers d'euros (Y) pour un échantillon de voitures d'occasion d'un même type:

X_1	X_2	Y
5	92	7.8
4	64	9.5
6	124	6.4
5	97	7.5
5	79	8.1
5	76	9.0
6	93	6.1
6	63	8.7
2	13	15.4
7	111	6.4
7	143	4.4

Si l'on veut étudier la liaison entre les variables potentiellement explicatives et la variable à expliquer (le prix Y), on peut commencer par visualiser les nuages de points (X_1, Y) et (X_2, Y):

```
donnees = [  
5 92 7.8; 4 64 9.5; 6 124 6.4; 5 97 7.5; 5 79 8.1; 5 76 9.0;  
6 93 6.1; 6 63 8.7; 2 13 15.4; 7 111 6.4; 7 143 4.4];  
// definition des variables  
X1=donnees(:,1); X2=donnees(:,2); Y=donnees(:,3);  
// allures des nuages  
xbasec();  
subplot(121), plot2d(X1,Y,-2,"111","Age",[1,0,8,16]);  
subplot(122), plot2d(X2,Y,-3,"111","Kms",[0,0,150,16]);
```

Question 11 *Que suggèrent les deux nuages?*

3.2 Régression linéaire simple

On propose tout d'abord le modèle $M1$:

$$Y = \beta + \alpha_1 X_1 + \varepsilon,$$

où ε est un échantillon de $\mathcal{N}(0, \sigma^2)$.

Question 12 *L'âge a-t-il une influence significative sur le prix?*

La fonction regression (`regression(M,Y,alpha)`) calcule, à partir de la donnée du modèle sous la forme $Y = M\theta + \varepsilon$, où Y est la variable à expliquer et M la matrice définissant le modèle, les éléments usuels:

- L'estimateur du paramètre θ : $\text{theta} = (M^t M)^{-1} M^t Y$.
- L'estimateur de σ^2 : $\text{sigma2} = MSE$.
- La table d'analyse de la variance pour l'hypothèse nulle "aucun régresseur n'est significatif".
- La valeur critique associée au niveau α .

Cliquer sur le lien ci-dessus pour obtenir le code de la fonction, le sauvegarder sous le nom `regression.sce` dans le répertoire où vous utilisez scilab. Pour charger la fonction, utiliser la commande: `getf 'regression.sce'`. Pour appliquer la fonction, utiliser la commande: `regression(M,Y,alpha)`

Pour le modèle, où on explique le prix en fonction de l'âge, on utilise les commandes suivantes:

```
// MODELE (M1) Y = theta1(0) + theta1(1) X1 + eps
M1 = [ones(X1) X1]; // matrice des regresseurs
// Pour obtenir les estimations de theta, sigma2,
// et la p-valeur
[theta1,sigma21,pval1]=regression(M1,Y,alpha);
```

On représente enfin la droite de régression sur le nuage:

```
xbasc();
// trace de la droite
z=1:0.1:8; t=theta1(1)+theta1(2)*z; xbasc();
plot2d(X1,Y,-2,"111","Age",[1,0,8,16]);
plot2d(z,t,[1],"000");
```

3.3 Régression linéaire multiple

On propose d'essayer à présent le modèle "complet"

$$Y = \gamma + \alpha_1 X_1 + \alpha_2 X_2 + \varepsilon.$$

Question 13 *Donner, à l'aide de la fonction `regression` des estimations sans biais des paramètres $(\gamma, \alpha_1, \alpha_2)$ et σ^2 .*

Question 14 *Test de l'utilité d'un régresseur: Tester, au niveau 5%, l'hypothèse nulle "le kilométrage n'a pas d'effet sur le prix" contre "c'est faux".*

La fonction `regression` n'effectue pas le test demandé (quel test effectue-t-elle?). Ecrire les commandes Scilab permettant de calculer la statistique et la p -valeur du test demandé, et conclure.

Question 15 *A l'aide de la question précédente, déterminer quel sous-modèle, parmi les deux sous-modèles à un seul régresseur, explique le mieux le prix?*