

# Tests de comparaison pour l'augmentation du volume de précipitation

October 10, 2017

## Contents

<b>1</b>	<b>Peut-on augmenter la quantité de pluie?</b>	<b>1</b>
<b>2</b>	<b>Un test de comparaison non paramétrique</b>	<b>2</b>
<b>3</b>	<b>Modèle gaussien</b>	<b>7</b>
3.1	Comparaison des moyennes . . . . .	8
3.2	(FACULTATIF) COMPARAISON DES VARIANCES DES 2 ÉCHANTILLONS GAUSSIENS	10
<b>4</b>	<b>Et si on testait si <math>X_i</math> et <math>Y_j</math> ont même loi</b>	<b>10</b>
4.1	Un peu de théorie . . . . .	11
4.2	La pratique . . . . .	11
4.3	En Scilab . . . . .	11

## 1 Peut-on augmenter la quantité de pluie?

Chacun sait que l'eau gèle en dessous de  $0^\circ C$ . Ce que l'on sait moins c'est que la phase liquide peut exister en-dessous de cette limite. Quand elle reste liquide en dessous de  $0^\circ C$  on dit que l'eau est surfondue. Ainsi les nuages sont presque toujours constitués de gouttelettes d'eau surfondue coexistant avec quelques cristaux de glace. C'est un état d'équilibre instable et tout apport supplémentaire de cristaux de glace ou d'iodure d'argent provoque l'évaporation rapide des gouttelettes en surfusion. La vapeur ainsi libérée va aussitôt se solidifier sur les cristaux de glace qui grossissent et précipitent vers le sol en se réchauffant. En effet, aux températures négatives, l'air atmosphérique est saturé de vapeur d'eau par rapport à la glace avant de l'être par rapport à l'eau surfondue. C'est cette propriété qui est exploitée pour forcer les précipitations. Ces pluies artificiellement provoquées peuvent être intéressantes pour l'agriculture comme en témoignent les nombreuses expériences d'ensemencement réalisées dans le monde au cours des années 1950 à 1970. C'est ainsi

qu'entre le 1er juin et le 23 août 1975 (83 jours), une expérience visant à accroître le volume des précipitations s'est tenue en Floride (USA). Les données suivantes sont extraites de l'article de W.L. Woodley, J. Simpson, R. Biondini et J. Berkeley (1977) : Rainfall Results 1970-75 : Florida Area Cumulus Experiment, publié dans Science, vol. 195, pp. 735-742.

Données d'ensemencement de nuages par iodure d'argent

Expérience $i$	1	2	3	4	5	6	7	8
Date $T_i$	0	1	3	4	6	9	18	25
Ensemencement $E_i$	0	1	1	0	1	0	0	0
Pluies $Z_i$	12.85	5.52	6.29	6.11	2.45	3.61	0.47	4.56
Expérience $i$	9	10	11	12	13	14	15	16
Date $T_i$	27	8	29	32	33	35	38	39
Ensemencement $E_i$	0	1	1	1	0	1	1	0
Pluies $Z_i$	6.35	5.06	2.76	4.05	5.74	4.84	11.86	4.45
Expérience $i$	17	18	19	20	21	22	23	24
Date $T_i$	53	55	56	59	65	68	82	83
Ensemencement $E_i$	0	1	0	1	1	0	1	0
Pluies $Z_i$	3.66	4.22	1.16	5.45	2.02	0.82	1.09	0.28

$T=[0, 1, 3, 4, 6, 9, 18, 25, 27, 8, 29, 32, 33, 35, 38, 39, 53, 55, 56, 59, 65, 68, 82, 83]$  ;  
 $E=[0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0]$  ;  
 $Z=[12.85, 5.52, 6.29, 6.11, 2.45, 3.61, 0.47, 4.56, 6.35, 5.06, 2.76, 4.05, 5.74]$  ;  
 $Z=[Z, 4.84, 11.86, 4.45, 3.66, 4.22, 1.16, 5.45, 2.02, 0.82, 1.09, 0.28]$  ;

Les sentiments mitigés des agriculteurs peuvent se résumer ainsi :

- le procédé est cher ;
- jusqu'à ce jour nous nous en sommes passés et nous pouvons continuer à nous en passer ;
- cependant si l'ensemencement de nuages par iodure d'argent est efficace, le procédé permettrait d'augmenter la production.

**Question 1** *Qu'attendent les paysans du statisticien?*

Les paragraphes qui suivent ont pour but de fournir la réponse du statisticien à l'aide d'une approche non-paramétrique (voir le livre de J. Bernier, E. Parent, JJ. Boreux, 2000, Stat pour l'environnement, Tec et Doc, où cet exemple est traité dans un cadre paramétrique).

## 2 Un test de comparaison non paramétrique

On note  $X_1, \dots, X_{n_1}$  les hauteurs de pluies avec ensemencement, et  $Y_1, \dots, Y_{n_2}$  les hauteurs de pluies sans ensemencement.

$X=Z(E==1)$ ;

$Y=Z(E==0)$ ;

On suppose que les variables aléatoires  $Y_1, \dots, Y_{n_2}$  sont indépendantes et de même fonction de répartition  $F$ . On suppose que les variables aléatoires  $X_1, \dots, X_{n_1}$  sont indépendantes et de même fonction de répartition  $F_\Delta = F(\cdot - \Delta)$ . Le but est donc de tester l'hypothèse  $H'_0 = \{\Delta \leq 0\}$ , ou plus exactement  $H_0 = \{\Delta = 0\}$ , contre son alternative  $H_1 = \{\Delta > 0\}$ . Remarquons que l'on ne fait pas d'hypothèse sur la loi de  $Y_i$ . Il existe beaucoup de tests pour ce genre de problème. On se propose d'utiliser le test de Mann et Whitney (1947) qui repose sur les rangs de l'échantillon  $X$  dans l'échantillon des deux populations  $(X, Y)$ .

On définit

$$U = \sum_{1 \leq i \leq n_1, 1 \leq j \leq n_2} \mathbf{1}_{\{Y_j < X_i\}} \quad \text{et} \quad W = \sum_{i=1}^{n_1} R_i,$$

où  $R_i$  est le rang de  $X_i$  dans le réordonnement croissant de la population totale  $(X, Y)$ . On rappelle la relation entre  $W$  et  $U$  :

$$U = W - \frac{n_1(n_1 + 1)}{2}.$$

De plus les lois sous  $H_0$  de  $U$  et  $W$  ne dépendent pas de la fonction de répartition  $F$ .

**Question 2** Quelles sont les valeurs maximales et minimales de  $W$  et  $U$ .

Sous  $H_0$ , la loi de  $(R_1, \dots, R_{n_1})$  est la loi de  $(\sigma_1, \dots, \sigma_{n_1})$ , où la permutation aléatoire  $\sigma = (\sigma_1, \dots, \sigma_{n_1}, \sigma_{n_1+1}, \dots, \sigma_{n_1+n_2})$  suit la loi uniforme sur l'ensemble des permutations de  $\{1, \dots, n_1 + n_2\}$ .

Il existe des tables pour la loi de  $W$ . Toutefois, il est aussi rapide d'en donner une approximation par simulation. On peut par exemple donner une estimation de  $\mathbb{P}(W > w)$  à l'aide de simulations de  $N$  variables aléatoires indépendantes de loi  $W$ . On commence par simuler la loi de  $W$  :

```
//n1 longueur de la population X
//n2 longueur de la population Y
//N nombre de simulations
fonction[W]=simulation_loi_W(n1,n2,N)
//longueur de la population totale
n=n1+n2;
vect=[1:n]';
//matrice dont chaque colonne est une permutation de {1,...,n}
```

```

R=grand(N,'prm',vect);
//on prend les n1 premiers éléments de chaque permutation
R=R([1:n1],:);
W=sum(R,'r');
endfunction;

```

On fait  $N = 10\,000$  simulations de  $W$ ,  $W=\text{simulation\_loi\_W}(n_1,n_2,10000)$ ; et on trace un histogramme (graphique 1) des valeurs obtenues pour visualiser une approximation de la loi de  $W$  :

```

xbasc();
histplot([0:2:n1*n2+n1*(n1+1)/2],W);

```

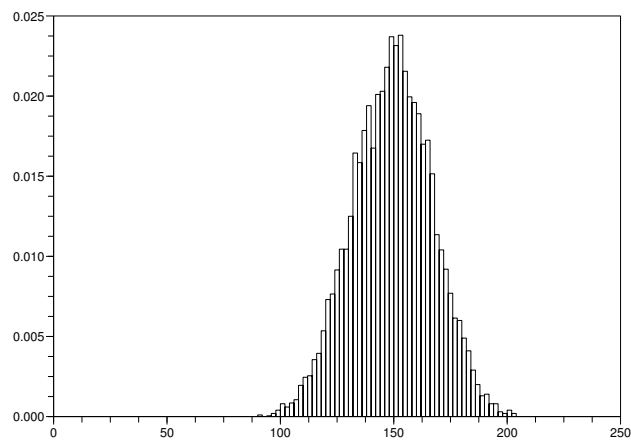


Figure 1: Simulation de la loi de  $W$

De même on simule la loi de  $U$  en se basant sur la relation entre  $U$  et  $W$  :

```

function[U]=simulation_loi_U(n1,n2,N)
W=simulation_loi_W(n1,n2,N);
U=W-0.5*n1*(n1+1);
endfunction;

```

On rappelle que sous  $H_0$ , on a

$$\mathbb{E}[U] = \frac{1}{2} n_1 n_2 \quad \text{et} \quad \text{Var}(U) = \frac{1}{12} n_1 n_2 (n_1 + n_2 + 1).$$

De plus :

$$\lim_{\min(n_1, n_2) \rightarrow \infty} \frac{U - \frac{1}{2} n_1 n_2}{\sqrt{\frac{1}{12} n_1 n_2 (n_1 + n_2 + 1)}} \stackrel{\text{loi}}{=} \mathcal{N}(0, 1).$$

On trace la densité gaussienne sur l'histogramme de l'approximation de la loi de  $U$  afin de visualiser la convergence en loi ci-dessus. Pour obtenir le graphique 2, on utilise la fonction Scilab

```
xbasc(); //nettoie la fenetre graphique
comparaison_U_Normale(n1,n2,N);
```

avec par exemple  $N = 10\,000$  simulations de  $U$ .

```
function comparaison_U_Normale(n1,n2,N)
U=simulation_loi_U(n1,n2,N);
t=(U-0.5*n1*n2)/sqrt(n1*n2*(n1+n2+1)/12);
x=[-4:2/sqrt(n1*n2*(n1+n2+1)/12):4];
histplot(x,t);
z=[-4:0.1:4];
y=exp(-z.^2/2)/sqrt(2*%pi);
plot2d(z,y);
endfunction;
```

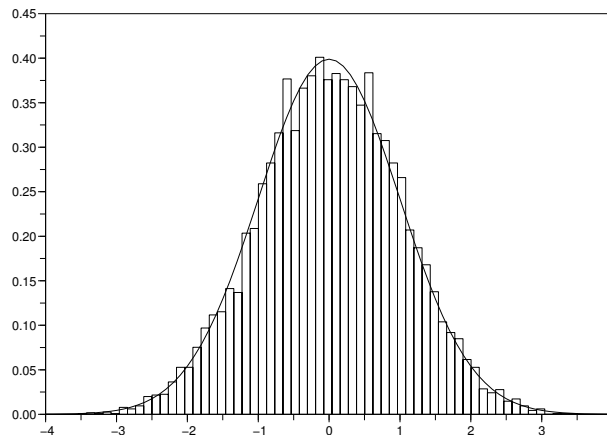


Figure 2: Comparaison entre la loi de  $U$  recentrée et renormalisée et une gaussienne  $\mathcal{N}(0, 1)$ .

Sous  $H_1$ , en revanche, les variables aléatoires  $X_i$  prennent en moyenne des valeurs plus grandes que les variables aléatoires  $Y_i$ . Bien que la loi de  $U$  sous  $H_1$  ne soit pas accessible,

car elle dépend de  $F$ , on remarque que  $W$  prend en moyenne des valeurs plus grandes sous  $H_1$  que sous  $H_0$ . En particulier, on peut montrer que sous  $H_1$

$$\lim_{\min(n_1, n_2) \rightarrow \infty} \frac{U - \frac{1}{2} n_1 n_2}{\sqrt{\frac{1}{12} n_1 n_2 (n_1 + n_2 + 1)}} \stackrel{\text{p.s.}}{=} +\infty.$$

**Question 3** Construire un test convergent à partir de la statistique  $U$  pour tester  $H_0$  contre  $H_1$ . Déterminer la région critique associée à ce test de niveau  $\alpha = 5\%$ . On utilisera l'approche asymptotique et l'approche directe. Dans ce dernier cas, comme la fonction de répartition de  $U$  n'est pas explicite, on utilisera la fonction `cdfUX` qui fournit par simulation les quantiles de la loi de  $U$ .

Cette fonction calcule  $U$  à partir de deux échantillons  $X$  et  $Y$  (elle ne tient pas compte des ex-aequo).

```
//calcul de U
function[U]=Mann_Whitney(X,Y)
n1=length(X);
//classe est la population totale (X,Y) triée par ordre croissant
//permut donne les rangs des éléments du vecteur classe dans
//la population (X,Y)
[classe,permut]=gsort([X Y], 'g', 'i');
//rangs stocke les rangs des éléments de X
//dans la population triée
rangs=find(permut<=n1);
//on fait la somme de ces rangs, cela donne W
W=sum(rangs);
//on calcule U
U=W-n1*(n1+1)/2;
endfunction;
```

Voici la fonction qui calcule par simulation les quantiles de la loi de  $U$  :

```
//renvoie x le quantile d'ordre p de la loi U : p=P(U<x)
//IC intervalle de confiance à 5% sur x
//N nombre de simulations pour estimer x
//n1 et n2 paramètres de U
function[x,IC]=cdfUX(n1,n2,p,N)
//simule N va de loi U
U=simulation_loi_U(n1,n2,N);
//tri par ordre croissant
[s,k]=gsort(U, 'g', 'i');
//estimateur
```

```

x=s(floor(p*N));
inf_IC=s(floor(p*N-sqrt(N)*1.96*sqrt(p*(1-p))));
sup_IC=s(floor(p*N+sqrt(N)*1.96*sqrt(p*(1-p))));
IC=[inf_IC,sup_IC];
endfunction;

```

**Question 4** Calculer la  $p$ -valeur associée au test précédent pour les données du problème, à l'aide de la fonction Scilab ci-dessous `cdfU`. Conclusion.

La fonction `[P,Q,IC_Q]=cdfU(u,n1,n2,N)` renvoie une estimation par simulation de  $P = \mathbb{P}(U \leq u)$ ,  $Q = 1 - P$ , ainsi que  $IC_Q$  l'intervalle de confiance asymptotique sur  $Q$  à 95%. On prendra  $N = 10\,000$ .

```

//n1 et n2 cardinaux des échantillons
//N nbombre de simulations
//P=P(U<=u)
//Q=1-P
//IC_Q intervalle de confiance sur Q
function[P,Q,IC_Q]=cdfU(u,n1,n2,N)
U=simulation_loi_U(n1,n2,N);
u=u*ones(U);
P=sum(U<=u)/N;
Q=sum(U>u)/N;
IC_Q=[Q-1.96*sqrt(Q*(1-Q))/sqrt(N),Q+1.96*sqrt(Q*(1-Q))/sqrt(N)];
endfunction;

```

**Question 5** Calculer la  $p$ -valeur pour les données du problème, à l'aide de l'approximation gaussienne. Conclusion.

Remarquons que ne pas rejeter  $H_0$  ne signifie pas que  $H_0$  soit vraie. Ainsi si on considère  $H'_0 = \{\Delta \leq 0\}$ , au lieu de  $H_0$ , on obtient la même région critique pour la statistique de test  $U$ . En particulier on ne peut pas distinguer  $H_0$  de  $H'_0$  et en fait même  $H_0$  de  $H''_0 = \{\mathbb{P}(X_1 \geq x) \leq \mathbb{P}(Y_1 \geq x) \quad \forall x \in \mathbb{R}\}$  ou de  $H'''_0 = \{\mathbb{P}(X_1 \leq Y_1) \leq 1/2\}$ .

### 3 Modèle gaussien

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.

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?

On peut répondre au point 1. à l'aide du test de Kolmogorov quand la moyenne et la variance sont connues a priori. Sinon, plusieurs tests de normalité sont détaillés dans la littérature. Le point 2. est abordé dans le paragraphe facultatif 3.2.

Nous admettons donc dans un premier temps ces deux hypothèses.

### 3.1 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

$$S_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2 = \frac{n_1}{n_1 - 1} \left[ \frac{1}{n_1} \sum_{i=1}^{n_1} X_i^2 - (\bar{X})^2 \right],$$

$$S_Y^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 = \frac{n_2}{n_2 - 1} \left[ \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i^2 - (\bar{Y})^2 \right].$$

```
// 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  $\chi^2$  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 6** Construire, à partir de la statistique  $T$ , un test d'égalité des moyennes en déterminant sa région critique, la valeur critique au niveau  $\alpha$  et la  $p$ -valeur.

**Question 7** 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` (`[Tobs,vc,pval]=ttest2(n1,Xbar,SX2,n2,Ybar,SY2,alpha)`) retourne la valeur de la statistique  $T$  observée, `Tobs`, la valeur critique associée au niveau  $\alpha$ , `vc`, et la  $p$ -valeur, `pval`, 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:

```
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 = cdfT("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 = cdfT("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 `cdft`, on peut consulter le manuel en utilisant la commande `help cdft`.

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 8

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 9** Reprendre les questions 6 et 7, si l'hypothèse nulle est  $\{\mu_1 \leq \mu_2\}$ .

## 3.2 (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 10 (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 11 (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 = cdff("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 de Fisher de paramètres `(k1, k2)`.

- `[p, q] = cdf("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`.

## 4 Et si on testait si $X_i$ et $Y_j$ ont même loi

Ce test est une variante du test de Kolmogorov Smirnov pour deux échantillons indépendants.

### 4.1 Un peu de théorie

On dispose aussi dans ce cas d'un test de Kolmogorov Smirnov. Il s'agit de comparer les fonctions de répartition empirique

$$FX_{n_1}(x) = \frac{1}{n_1} \sum_{i=1}^{n_1} 1_{\{X_i \leq x\}} \text{ et } FY_{n_2}(x) = \frac{1}{n_2} \sum_{i=1}^{n_2} 1_{\{Y_i \leq x\}}.$$

On définit la statistique du test

$$\zeta_{n_1, n_2} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \sup_{x \in \mathbb{R}} |FX_{n_1}(x) - FY_{n_2}(x)|.$$

Si l'hypothèse nulle  $H_0 = \{Loi(X_i) = Loi(Y_i)\}$  est vraie alors  $\zeta_{n_1, n_2}$  tend en loi vers une variable aléatoire  $S$  dont la loi est indépendante de celles de  $X$  et  $Y$ . De plus on connaît sa fonction de répartition

$$K(s) = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2 s^2} \quad s > 0.$$

Si l'hypothèse nulle est fautive alors  $\zeta_{n_1, n_2}$  diverge presque sûrement.

### 4.2 La pratique

Utilisation de ce résultat :

- on dispose de l'observation des échantillons  $X_1 = x_1, \dots, X_{n_1} = x_{n_1}$  et  $Y_1 = y_1, \dots, Y_{n_2} = y_{n_2}$
- on calcule les fonctions de répartition empiriques  $FX_{n_1}$  et  $FY_{n_2}$  à partir des échantillons
- on calcule la valeur de  $\zeta_{n_1, n_2} = z_{n_1, n_2}$
- on calcule la p-valeur  $\alpha_{n_1, n_2} = \mathbb{P}(S \geq z_{n_1, n_2})$ .
- si  $\alpha_{n_1, n_2}$  prend des valeurs faibles *inférieures* à  $\alpha = 5\%$  typiquement alors on rejette l'hypothèse  $H_0$  sinon on l'accepte.
- $\alpha$  représente le niveau de confiance du test.

### 4.3 En Scilab

La fonction `ks2` (`[D,q]=ks2(X,Y)`) permet de réaliser le test précédemment décrit. Pour utiliser cette fonction, cliquer sur le lien ci-dessus et enregistrer el fichier sous le nom `ks2.sce` dans le répertoire dans lequel vous travaillez. Pour charger la fonction, utiliser la commande : `exec 'ks2.sce'`.

Explications des paramètres :

- `x` : vecteur ligne des observations du premier échantillon
- `y` : vecteur ligne des observations du second échantillon

Cette fonction renvoie :

1.  $D : \zeta_{n_1, n_2} / \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$
2.  $q : \mathbb{P}\left(S \geq \sqrt{\frac{n_1 n_2}{n_1 + n_2}} D\right) = \mathbb{P}(S \geq \zeta_{n_1, n_2})$ .

**Question 12** *Le procédé d'ensemencement proposé change-t-il significativement la loi des hauteurs de pluie?*