TD 9 : tests de Kolmogorov-Smirnov

Les questions marquées d'un astérisque **(*)** *sont facultatives.*

Pour les exercices faisant intervenir un test de Kolmogorov-Smirnov, on se réfèrera aux tables de quantiles à la fin de la feuille, tirées de "Advanced Statistics from an Elementary Point of View", par Michael J. Panik. Attention à bien utiliser le α correspondant au Two-Sided Test !

Pour information, il est possible de réaliser des tests de Kolmogorov-Smirnov dans R grâce à la fonction `ks.test`*. Les tests du χ^2 se réalisent avec* `chisq.test`*, de Student avec* `t.test`*.*

Exercice 1. (Fonction de répartition) Soit X une variable aléatoire à valeurs dans $\mathbb{R}$.

1. Rappeler la définition de la fonction de répartition F_X de X. La fonction F_X est-elle injective ? Surjective ? Bijective ?
2. Rappeler la définition de l'inverse généralisée F_X^{-1} de F_X.
3. A-t-on (si oui, le montrer, si non, donner un contre-exemple et **(*)** corriger l'énoncé) :
   (a) Pour tout $x \in \mathbb{R}$, $(F_X^{-1} \circ F_X)(x) = x$?
   (b) Pour tout $q \in [0,1]$, $(F_X \circ F_X^{-1})(q) = q$?
4. Soit $U \sim \mathcal{U}([0,1])$. Montrer que $F_X^{-1}(U)$ a la même loi que X.
   On utilisera que si deux variables X et Y vérifient : pour tout $t \in \mathbb{R}$, $\mathbb{P}(X \leqslant t) = \mathbb{P}(Y \leqslant t)$, alors elles ont la même loi.
5. Supposons que la loi de X n'a pas d'atomes. Montrer que $F_X(X) \sim \mathcal{U}([0,1])$. **(*)** Et si la loi de X a un atome ?

Exercice 2. On observe l'échantillon de 20 observations (supposées indépendantes), trié par ordre croissant par souci de lisibilité, suivant :

0.278	0.452	0.464	0.494	0.496	0.505	0.576	0.592	0.602	0.608
0.661	0.683	0.690	0.696	0.704	0.728	0.754	0.850	0.902	0.949

On souhaite savoir s'il a été généré suivant la loi uniforme sur $[0,1]$.

1. Formuler les hypothèses H0 et H1.
2. Quelle est la fonction de répartition F_U de la loi uniforme ?
3. Rappeler la définition de la fonction de répartition empirique F_n fondée sur un n-échantillon $(X_1, \dots, X_n)$. On note $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$ les statistiques d'ordre de l'échantillon, c'est-à-dire l'échantillon trié par ordre croissant.
4. Énoncer le théorème de Glivenko-Cantelli.
5. Vers quelle loi convergerait $\sqrt{n}\|F_n - F_U\|_\infty$ si les observations suivaient la loi uniforme ?
6. Si les observations suivaient une loi de fonction de répartition $F^\star \neq F_U$, vers quoi convergerait $\|F_n - F_U\|_\infty$? Et $\sqrt{n}\|F_n - F_U\|_\infty$?
7. En déduire une région de rejet du test ayant pour statistique de test $\|F_n - F_U\|_\infty$.
8. Fournir les bornes de cette région de rejet en se fondant sur les tables de quantiles à la fin de l'énoncé.
   Attention, deux points peuvent vous induire en erreur :
   (a) *Référez-vous à la ligne « Two-Sided Test » pour le niveau des quantiles ;*
   (b) *La table est construite pour la statistique $\|F_n - F_U\|_\infty$, pas $\sqrt{n}\|F_n - F_U\|_\infty$.*

9. Calculer $\|F_n - F_U\|_\infty$.
10. Conclure.
11. Il est également possible de faire un test du chi-deux d'adéquation. Pour cela, on compte le nombre d'observations tombant dans des intervalles de poids similaire. Ici, on choisit $[0, \frac{1}{3}]$, $[\frac{1}{3}, \frac{2}{3}]$ et $[\frac{2}{3}, 1]$. Effectuer ce test et comparer les conclusions.

Exercice 3. (Test à deux échantillons) On observe deux échantillons, triés par ordre croissant :

$$-0.367 \quad 0.109 \quad 0.424 \quad 0.517 \quad 3.405$$

et

$$0.396 \quad 0.783 \quad 0.810 \quad 1.290 \quad 1.604 \quad 1.640 \quad 1.799 \quad 1.857 \quad 2.579 \quad 2.787$$

Testez s'ils ont la même loi.

La table est construite pour la statistique $\|F_n - G_m\|_\infty$ lorsque F_n et G_m sont des fonctions des répartition empiriques de deux échantillons indépendants, le premier composé de n variables i.i.d. uniformes sur $[0, 1]$ et le second composé de m variables i.i.d. uniformes sur $[0, 1]$. On justifiera soigneusement pourquoi il est possible d'utiliser cette table de quantiles dans cet exercice.

Exercice 4. (*) (Correction de Lilliefors) Soit $(X_1, \dots, X_n)$ un échantillon de variables aléatoires i.i.d. On souhaite tester si elles suivent une loi normale, mais sans connaître leur moyenne ni leur variance : il n'est donc pas possible d'utiliser le test de Kolmogorov-Smirnov.

Soit $\bar{X}_n$ la moyenne empirique de l'échantillon et $\hat{\sigma}_n$ son écart-type empirique. Pour tout $i \in \{1, \dots, n\}$, on note $Z_i = (X_i - \bar{X}_n)/\hat{\sigma}_n$ l'observation X_i recentrée et renormalisée. Soit F la fonction de répartition de la loi normale centrée réduite et F_n^Z la fonction de répartition empirique de l'échantillon $(Z_i)_{1 \leqslant i \leqslant n}$. La statistique du test de Lilliefors est

$$D = \|F_n^Z - F\|_\infty.$$

1. En notant $Z_{(1)} \leqslant \cdots \leqslant Z_{(n)}$ les statistiques d'ordre de l'échantillon $(Z_i)_{1 \leqslant i \leqslant n}$, montrer que
$$D = \max_{1 \leqslant i \leqslant n} \left(F(Z_{(i)}) - \frac{i-1}{n}, \frac{i}{n} - F(Z_{(i)}) \right).$$
2. Montrer que la loi de D ne dépend que de la taille de l'échantillon et non de sa moyenne ou de sa variance.
3. En déduire un test de normalité de l'échantillon $(X_1, \dots, X_n)$.
4. **(*)** La loi des Z_i ne dépend ni de la moyenne, ni de la variance des X_i. Pourquoi ne peut-on pas leur appliquer le test de Kolmogorov-Smirnov ? (Cela consisterait à remplacer F, fonction de répartition de la loi normale centrée réduite, par la fonction de répartition de Z dans la définition de D.)

Pour en savoir plus sur les tests de normalité, vous pouvez vous référer au cours `http://eric.univ-lyon2.fr/~ricco/cours/cours/Test_Normalite.pdf`

Table A.14 Quantiles of the Kolmogorov-Smirnov Test Statistics D_n

The table gives the upper $100(1-\alpha)\%$ quantile $\hat{d}_{n,1-\alpha}$ of the sampling distribution of $\hat{D}_n$ such that $P(\hat{D}_n \leq \hat{d}_{n,1-\alpha}) = 1-\alpha$ or $P(\hat{D}_n \geq \hat{d}_{n,1-\alpha}) = \alpha$ (e.g., for $n = 20$ and $\alpha = 0.05$, the one-tail critical region is $\mathcal{R} = \{\hat{d}_{20} | \hat{d}_{20} \geq \hat{d}_{20,0.95} = 0.265\}$; the two-tail critical region is $\mathcal{R} = \{\hat{d}_{20} | \hat{d}_{20} \geq \hat{d}_{20,0.95} = 0.294\}$).

One-Sided Test $1-\alpha =$	**0.90**	**0.95**	**0.975**	**0.99**	**0.995**	$1-\alpha =$	**0.90**	**0.95**	**0.975**	**0.99**	**0.995**
Two-Sided Test $1-\alpha =$	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**	$1-\alpha =$	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**
$n = 1$	0.900	0.950	0.975	0.990	0.995	$n = 21$	0.226	0.259	0.287	0.321	0.344
2	0.684	0.776	0.842	0.900	0.929	**22**	0.221	0.253	0.281	0.314	0.337
3	0.565	0.636	0.708	0.785	0.829	**23**	0.216	0.247	0.275	0.307	0.330
4	0.493	0.565	0.624	0.689	0.734	**24**	0.212	0.242	0.269	0.301	0.323
5	0.447	0.509	0.563	0.627	0.669	**25**	0.208	0.238	0.264	0.295	0.317
6	0.410	0.468	0.519	0.577	0.617	**26**	0.204	0.233	0.259	0.290	0.311
7	0.381	0.436	0.483	0.538	0.576	**27**	0.200	0.229	0.254	0.284	0.305
8	0.358	0.410	0.454	0.507	0.542	**28**	0.197	0.225	0.250	0.279	0.300
9	0.339	0.387	0.430	0.480	0.513	**29**	0.193	0.221	0.246	0.275	0.295
10	0.323	0.369	0.409	0.457	0.489	**30**	0.190	0.218	0.242	0.270	0.290
11	0.308	0.352	0.391	0.437	0.468	**31**	0.187	0.214	0.238	0.266	0.285
12	0.296	0.338	0.375	0.419	0.449	**32**	0.184	0.211	0.234	0.262	0.281
13	0.285	0.325	0.361	0.404	0.432	**33**	0.182	0.208	0.231	0.258	0.277
14	0.275	0.314	0.349	0.390	0.418	**34**	0.179	0.205	0.227	0.254	0.273
15	0.266	0.304	0.338	0.377	0.404	**35**	0.177	0.202	0.224	0.251	0.269
16	0.258	0.295	0.327	0.366	0.392	**36**	0.174	0.199	0.221	0.247	0.265
17	0.250	0.286	0.318	0.355	0.381	**37**	0.172	0.196	0.218	0.244	0.262
18	0.244	0.279	0.309	0.346	0.371	**38**	0.170	0.194	0.215	0.241	0.258
19	0.237	0.271	0.301	0.337	0.361	**39**	0.168	0.191	0.213	0.238	0.255
20	0.232	0.265	0.294	0.329	0.352	**40**	0.165	0.189	0.210	0.235	0.252

Adapted from L.H. Miller, "Tables of Percentage Points of Kolmogorov Statistics," *JASA,* 51, 1956, 111–121. Reprinted with permission from *The Journal of the American Statistical Association*. Copyright [1956] by the American Statistical Association. All rights reserved.

Table A.15 Quantiles of the Kolmogorov-Smirnov Test Statistic $D_{n,m}$ When $n = m$

The table gives the upper $100(1-\alpha)$ % quantile $\hat{d}_{n,m}$ of the sampling distribution of $\hat{D}_{n,m}$ such that $P(\hat{D}_{n,m} \leq \hat{d}_{n,m,1-\alpha}) = 1-\alpha$ or $P(\hat{D}_{n,m} \geq \hat{d}_{n,m,1-\alpha}) = \alpha$ (e.g., for $n = m = 15$ and $\alpha = 0.05$, the one-tail critical region is $\mathcal{R} = \{\hat{d}_{15,15} | \hat{d}_{15,15} \geq \hat{d}_{15,15,0.95} = 0.40\}$; the two-tail critical region is $\mathcal{R} = \{\hat{d}_{15,15} | \hat{d}_{15,15} \geq \hat{d}_{15,15,0.95} = 0.467\}$).

One-Sided Test $1-\alpha =$	**0.90**	**0.95**	**0.975**	**0.99**	**0.995**	$1-\alpha =$	**0.90**	**0.95**	**0.975**	**0.99**	**0.995**
Two-Sided Test $1-\alpha =$	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**	$1-\alpha =$	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**
$n = 3$	2/3	2/3				$n = 20$	6/20	7/20	8/20	9/20	10/20
4	3/4	3/4	3/4			**21**	6/21	7/21	8/21	9/21	10/21
5	3/5	3/5	4/5	4/5	4/5	**22**	7/22	8/22	8/22	10/22	10/22
6	3/6	4/6	4/6	5/6	5/6	**23**	7/23	8/23	9/23	10/23	10/23
7	4/7	4/7	5/7	5/7	5/7	**24**	7/24	8/24	9/24	10/24	11/24
8	4/8	4/8	5/8	5/8	6/8	**25**	7/25	8/25	9/25	10/25	11/25
9	4/9	5/9	5/9	6/9	6/9	**26**	7/26	8/26	9/26	10/26	11/26
10	4/10	5/10	6/10	6/10	7/10	**27**	7/27	8/27	9/27	11/27	11/27
11	5/11	5/11	6/11	7/11	7/11	**28**	8/28	9/28	10/28	11/28	12/28
12	5/12	5/12	6/12	7/12	7/12	**29**	8/29	9/29	10/29	11/29	12/29
13	5/13	6/13	6/13	7/13	8/13	**30**	8/30	9/30	10/30	11/30	12/30
14	5/14	6/14	7/14	7/14	8/14	**31**	8/31	9/31	10/31	11/31	12/31
15	5/15	6/15	7/15	8/15	8/15	**32**	8/32	9/32	10/32	12/32	12/32
16	6/16	6/16	6/25	8/16	12/15	**34**	8/34	10/34	11/34	12/34	13/34
17	9/29	7/17	7/17	8/22	9/17	**36**	9/36	10/36	11/36	12/36	13/36
18	6/18	7/18	8/18	9/18	9/19	**38**	9/38	10/38	11/38	13/38	14/38
19	6/19	7/19	8/19	9/19	9/19	**40**	9/40	10/40	12/40	13/40	14/40
						Approximation for $n > 40$:	$\frac{1.52}{\sqrt{n}}$	$\frac{1.73}{\sqrt{n}}$	$\frac{1.92}{\sqrt{n}}$	$\frac{2.15}{\sqrt{n}}$	$\frac{2.30}{\sqrt{n}}$

Adapted from Z.W. Birnbaum and R.A. Hall, "Small Sample Distribution for Multisample Statistics of the Smirnov Type," *The Annals of Mathematical Statistics*, 31, 1960, 710–720, with kind permission from the Institute of Mathematical Statistics.

Table A.16 Quantiles of the Kolmogorov-Smirnov Test Statistic $D_{n,m}$ When $n \neq m^*$

The table gives the upper $100(1-\alpha)\%$ quantile $\hat{d}_{n,m}$ of the sampling distribution of $\hat{D}_{n,m}$ such that $P(\hat{D}_{n,m} \leq \hat{d}_{n,m,1-\alpha}) = 1 - \alpha$ or $P(\hat{D}_{n,m} \geq \hat{d}_{n,m,1-\alpha}) = \alpha$ (e.g., for $n = 6$, $m = 10$, and $\alpha = 0.05$, the one-tail critical region is $\mathcal{R} = \{\hat{d}_{6,10} | \hat{d}_{6,10} \geq \hat{d}_{6,10,0.95} = 0.567\}$; the two-tail critical region is $\mathcal{R} = \{\hat{d}_{6,10} | \hat{d}_{6,10} \geq \hat{d}_{6,10,0.95} = 0.633\}$).

One-Sided Test	$\mathbf{1-\alpha =}$	**0.90**	**0.95**	**0.975**	**0.99**	**0.995**
Two-Sided Test	$\mathbf{1-\alpha =}$	**0.80**	**0.90**	**0.950**	**0.98**	**0.990**
$n = 1$	$m = 9$	17/18				
	10	9/10				
$n = 2$	3	5/6				
	4	3/4				
	5	4/5	4/5			
	6	5/6	5/6			
	7	5/7	6/7			
	8	3/4	7/8	7/8		
	9	7/9	8/9	8/9		
	10	7/10	4/5	9/10		
$n = 3$	$m = 4$	3/4	3/4			
	5	2/3	4/5	4/5		
	6	2/3	2/3	5/6		
	7	2/3	5/7	6/7	6/7	
	8	5/8	3/4	3/4	7/8	
	9	2/3	2/3	7/9	8/9	8/9
	10	3/5	7/10	4/5	9/10	9/10
	12	7/12	2/3	3/4	5/6	11/12
$n = 4$	$m = 5$	3/5	3/4	4/5	4/5	
	6	7/12	2/3	3/4	5/6	5/6
	7	17/28	5/7	3/4	6/7	6/7
	8	5/8	5/8	3/4	7/8	7/8
	9	5/9	2/3	3/4	7/9	8/9
	10	11/20	13/20	7/10	4/5	4/5
	12	7/12	2/3	2/3	3/4	5/6
	16	9/16	5/8	11/16	3/4	13/16
$n = 5$	$m = 6$	3/5	2/3	2/3	5/6	5/6
	7	4/7	23/35	5/7	29/35	6/7
	8	11/20	5/8	27/40	4/5	4/5
	9	5/9	3/5	31/45	7/9	4/5
	10	1/2	3/5	7/10	7/10	4/5
	15	8/15	3/5	2/3	11/15	11/15
	20	1/2	11/20	3/5	7/10	3/4

Table A.16 (Contd.)

One-Sided Test	$1-\alpha =$	**0.90**	**0.95**	**0.975**	**0.99**	**0.995**
Two-Sided Test	$1-\alpha =$	**0.80**	**0.90**	**0.950**	**0.98**	**0.990**
$n = 6$	$m = 7$	23/42	4/7	29/42	5/7	5/6
	8	1/2	7/12	2/3	3/4	3/4
	9	1/2	5/9	2/3	13/18	7/9
	10	1/2	17/30	19/30	7/10	11/15
	12	1/2	7/12	7/12	2/3	3/4
	18	4/9	5/9	11/18	2/3	13/18
	24	11/24	1/2	7/12	5/8	2/3
$n = 7$	$m = 8$	27/56	33/56	5/8	41/56	3/4
	9	31/63	5/9	40/63	5/7	47/63
	10	33/70	39/70	43/70	7/10	5/7
	14	3/7	1/2	4/7	9/14	5/7
	28	3/7	13/28	15/28	17/28	9/14
$n = 8$	$m = 9$	4/9	13/24	5/8	2/3	3/4
	10	19/40	21/40	23/40	27/40	7/10
	12	11/24	1/2	7/12	5/8	2/3
	16	7/16	1/2	9/16	5/8	5/8
	32	13/32	7/16	1/2	9/16	19/32
$n = 9$	$m = 10$	7/15	1/2	26/45	2/3	31/45
	12	4/9	1/2	5/9	11/18	2/3
	15	19/45	22/45	8/15	3/5	29/45
	18	7/18	4/9	1/2	5/9	11/18
	36	13/36	5/12	17/36	19/36	5/9
$n = 10$	$m = 15$	2/5	7/15	1/2	17/30	19/30
	20	2/5	9/20	1/2	11/20	3/5
	40	7/20	2/5	9/20	1/2	
$n = 12$	$m = 15$	23/60	9/20	1/2	11/20	7/12
	16	3/8	7/16	23/48	13/24	7/12
	18	13/36	5/12	17/36	19/36	5/9
	20	11/30	5/12	7/15	31/60	17/30
$n = 15$	$m = 20$	7/20	2/5	13/30	29/60	31/60
$n = 16$	$m = 20$	27/80	31/80	17/40	19/40	41/80
Large-sample approximation		$1.07\sqrt{\frac{m+n}{mn}}$	$1.22\sqrt{\frac{m+n}{mn}}$	$1.36\sqrt{\frac{m+n}{mn}}$	$1.52\sqrt{\frac{m+n}{mn}}$	$1.63\sqrt{\frac{m+n}{mn}}$

*Let n be the smaller sample size and let m be the larger sample size. If this table does not cover n and m, use the large sample approximation.

Adapted from F.J. Massey, "Distribution Table for the Deviation Between Two Sample Cumulatives," *The Annals of Mathematical Statistics*, 23, 1952, 435–441. Corrections appear in Davis, L.S. (1958), Mathematical Tables and other Aids to Computation, 12, 1952, 262–263, with kind permission from the Institute of Mathematical Statistics.