TD2 Tests et intervalles de confiance avec R – Master 1 ETEC

Exercice 1. On a mesuré le poids de raisin par souche sur 10 souches prises au hasard dans une vigne. On a obtenu les résultats suivants (en kg) :

2,4 ; 3,2 ; 3,6 ; 4,1 ; 4,3 ; 4,7 ; 5,4 ; 5,9 ; 6,5 ; 6,9.

Déterminez une estimation ponctuelle non biaisée de la moyenne et de la variance de la population dont ces souches sont extraites. (*mean(), var()*)
Donnez un intervalle de confiance de la moyenne au risque de 5% en supposant que le poids de raisin par souche suit une loi normale au niveau de la vigne (*t-test*).
Donner au risque de 5% un intervalle de confiance de la variance, puis de l'écart type de la population (*programmer par bootstrap cf web*).

Exercice 2. Les données suivantes ont été obtenues sur des échantillons d'individus d'une région d'Europe. Le caractère étudié est le poids du cerveau exprimé en grammes pour des sujets adultes. On suppose que cette variable suit une loi normale dans les deux cas.

Hommes

Centre de classe	1170	1220	1270	1320	1370	1420	1470	Total
Effectif	5	36	45	50	61	49	17	263

Femmes

Centre de classe	1070	1120	1170	1220	1270	1320	1370	Total
Effectif	12	22	45	54	52	20	10	215

Déterminer un intervalle de confiance au risque de 5% pour la moyenne de la population des hommes, puis pour celle des femmes. *Penser en termes de moyennes et écart types pondérés. Package SDMTools ou à programmer?*

Exercice 3. Sur une parcelle de soja, on a mesuré la hauteur en cm de 100 plantes à l'âge de 6 semaines. Les résultats obtenus sont les suivants :

Hauteur en cm	36	37	38	39	40	41
Effectifs	6	11	26	32	14	11

Dans l'hypothèse d'une population gaussienne, déterminer un intervalle de confiance de la variance de la population, à $\alpha = 0.05$.

Exercice 4. Une entreprise de production de graines veut vérifier la faculté germinative d'une espèce, c'est-à-dire la probabilité *p* pour qu'une graine, prise au hasard dans la production germe. Sur un échantillon de 400 graines, on observe que 330 graines germent. Quel est l'intervalle de confiance de *p* au risque 5% ? au risque 1% ? (*binom.test()*)

Exercice 5. Dans la population française, le pourcentage d'individus dont le sang est de Rhésus négatif est de 15%. Dans un échantillon représentatif de 200 Basques français on observe que 44 personnes sont de Rhésus négatif. Peut-on dire, au risque 0.05, que les Basques diffèrent du reste de la France en ce qui concerne le caractère Rhésus ?

Exercice 6. Pour traiter un certain type de tumeur, on a utilisé deux schémas thérapeutiques.
Sur 40 malades traités selon le schéma A, on a observé une mortalité à 5 ans de 15%.
Sur 60 malades traités selon le schéma B, la mortalité à 5 ans a été de 25%.
Si l'on considère la mortalité à 5 ans, peut-on dire que les schémas A et B diffèrent significativement au risque 10% ? au risque 5% ? (*fisher.test() ou prop.test()*)

Exercice 7. Les spécifications d'un certain médicament indiquent que chaque comprimé doit contenir 2,5 g de substance active. 100 comprimés sont choisis au hasard dans la production, puis analysés. Ils contiennent en moyenne 2,6 g de substance active, avec un écart type estimé s=0,4 g. Peut-on dire que le médicament respecte les spécifications (α=0.05) ?

Exercice 8. Dans une étude sur les mécanismes de détoxication, Alary et Bouleau (2009) dosèrent, en microgramme par gramme, la concentration du DDT et de ses dérivés (DDD et DDE), contenus dans des brochets du Nord *(Esox lucius),* capturés dans la rivière Richelieu (prov. de Québec). Les données relatives aux brochets de 2 ans et de 3 ans sont présentées dans le tableau 1. On suppose que la différence suit une loi normale. Les moyennes de ces deux échantillons prélevés indépendamment l'un de l'autre diffèrent-elles de façon hautement significative? (*var.test() ou bartlett.test(), t.test(),*)

Tableau 1. **Concentration en DDT et ses dérivés chez le brochet du Nord**

2 ans concentration en DDT + DDD + DDE ($mg.kg^{-1}$)	3ans concentration en DDT + DDD + DDE ($mg.kg^{-1}$)
0,184	0,354
0,193	0,359
0,197	0,361
0,198	0,362
0,199	0,364
0,199	0,373
0,206	0,382
0,216	0,258
0,403	0,413

Exercice 9. Une compagnie d'assurances a décidé d'équiper ses bureaux de micro-ordinateurs. Elle désire acheter ces micro-ordinateurs à deux fournisseurs différents pour autant qu'il n'y ait pas de différence significative de fiabilité entre les deux marques. Elle teste un échantillon de 8 micro-ordinateurs de la marque 1 et un échantillon de 8 micro-ordinateurs de la marque 2, en relevant le temps écoulé (en heures) avant la première panne. Les données observées sont présentées ci-dessous. On suppose que la variable suit une loi normale. Question : Notez-vous une différence significative ?

Marque 1

2132	2275	2374	2400	2437	2402	2822	2892
2780	2833	2714	2705	2850	2799	2849	2742

Marque 2

2678	2823	2713	2786	2700	2831	2823	2779
2766	2773	2828	2769	2836	2715	2846	2708

Exercice 10. Dans une étude sur le traitement des eaux usées (Beak, 1993) l'efficacité de deux filtres, l'un en fibre de verre et l'autre en papier filtre Whatman n^0 40, a été testée. Sur des prélèvements de 200 millilitres d'eau provenant d'usines de pâtes à papier, la quantité de solides en suspension retenus par les deux filtres a été mesurée. Les résultats de ces analyses figurent au tableau 1. L'efficacité du filtre en fibre de verre est-elle supérieure à celle du papier filtre ? Les résultats sont supposés suivre une loi normale.

TABLEAU 1 Quantité de solides en suspension retenus par deux types de filtres

Numéro du prélèvement	Solides en suspension mg / L	Solides en suspension mg / L
	Filtre fibre de verre	Papier filtre
1	65	53
2	80	63
3	89	90
4	64	52
5	68	64
6	68	50
7	86	88
8	54	35
9	91	102
10	77	59

Exercice 11. Deux espèces d'ostracodes ont été prélevées dans du matériel provenant d'un forage. On ne sait rien de la loi suivie par la variable aléatoire étudiée au niveau des populations. Observe-t-on une différence dans l'occurrence de ces deux espèces au cours du temps ? On peut utiliser la profondeur comme mesure ordinale du temps, et utiliser une hypothèse nulle d'égalité des profondeurs médianes. (*wilcox.test()*)

Espèce A		Espèce B	
Profondeur	rang	Profondeur	Rang
242		202	
253		203	
271		208	
292		233	
305		251	
332		258	
335		271	
337		282	
338		283	
350		301	
357		308	
364		314	
365		327	
371			
372			
385			
401			
402			
410			
412			
418			
423			
427			

Exercice 12. Dans le cadre d'une expertise de validation d'un protocole d'analyse pétrologique, on analyse par une nouvelle méthode 12 roches et les résultats sont comparés à une méthode de référence. On ne sait rien de la loi suivie par la variable aléatoire étudiée au niveau des populations. Les résultats sont les suivants :

Roche n°	Nouvelle méthode	Méthode de référence
1	9,2	9,5
2	10	9
3	9	8,8
4	9,4	9,5
5	10,1	9,1
6	9,5	10
7	10	10,1
8	10,3	9,3
9	10,2	9
10	10,2	9,7
11	9,8	9,1
12	10,1	9,3

Exercice 13. Le tableau ci dessous donne un échantillon de 40 notes provenant d'un examen national. On ne sait rien de la loi suivie par la variable aléatoire étudiée au niveau des populations. Tester au risque 0,05 l'hypothèse que la note médiane pour les participants est (a) 66 et (b) 75.

71	67	55	64	82	66	74	58	79	61
78	46	84	93	72	54	78	86	48	52

Exercice 14. On dispose de deux informations sur une série de 9 faciès : la teneur en $CaCO_3$ (variable *x*) et la cohésion (variable *y*, codée de 0 à 20). On ne sait rien de la loi suivie par la variable aléatoire étudiée au niveau des populations. Calculez le coefficient r_s de Spearman. Est-il significatif ? (*cor.test()*)

Faciès	*x*	Rang	*y*
1	12		14
2	15		7
3	18		20
4	22		18
5	3		8
6	7		3
7	4		6
8	17		12
9	20		19

Exercice 15 : Comparer les échantillons suivants :

E1 :	3	9.8	2.0	5.2	3.6	5.9	8.5	9.4		
E2 :	9.3	12.5	11.3	7.6	3.2	8.6	7.2	14.2	9.6	3.2

On ne sait rien de la loi suivie par la variable aléatoire étudiée au niveau des populations.

Exercice 16 : On considère les classements en silice et en fer, d'un groupe de 12 roches :

Roches	A	B	C	D	E	F	G	H	I	J	K	L
Silice	6	4	12	1	10	5	8	2	11	7	3	9
Fer	3	9	11	2	12	4	10	5	8	1	6	7

On ne sait rien de la loi suivie par la variable aléatoire étudiée au niveau des populations. Y a-t-il une corrélation significative entre les rangs des deux paramètres classés ?

Exercice 17 : Au départ d'une course de chevaux, il y a habituellement huit positions de départ et la position numéro 1 est la plus proche de la palissade. On soupçonne qu'un cheval a plus de chances de gagner quand il porte un numéro faible, c'est-à-dire qu'il est plus proche de la palissade intérieure. Voici les données de 144 courses :

Numéro de départ	1 2 3 4 5 6 7 8
Nombre de victoires d'un cheval ayant ce numéro	29 19 18 25 17 10 15 11

Poser les hypothèses à tester (hypothèse nulle et hypothèse alternative). (*chisq.test()*)

Exercice 18 : Nous désirons déterminer si le taux de natalité peut être expliqué uniquement par le taux d'urbanisation. Il s'agit donc d'estimer le taux de natalité en fonction du taux d'urbanisation, à l'aide d'une droite de régression. On considère le modèle normal ici. (*lm()*)

a. Déterminer les meilleurs estimateurs de a et de b pour $y = ax + b$
b. Déterminer R^2

c. Quelle proportion de variation de la variable Y est expliquée par la relation linéaire ainsi déterminée ?
d. La corrélation est-elle significative ?
e. Donnez l'intervalle de confiance de a à 95%
f. Même chose que les questions d et e mais sur l'ordonnée à l'origine.

Pays	Taux de natalité	Taux d'urbanisation
Canada	16.2	55.0
Costa Rica	30.5	27.3
Cuba	16.9	33.3
États Unis	16.0	56.5
El Salvador	40.2	11.5
Guatemala	38.4	14.2
Haïti	41.3	13.9
Honduras	43.9	19.0
Jamaïque	28.3	33.1
Mexique	33.9	43.2
Nicaragua	44.2	28.5

Exercice 19. L'accélération d'un corps en chute libre peut se déduire de mesures successives de sa hauteur y_i à des temps t_i régulièrement espacés (évalués avec un stroboscope par exemple) après détermination du meilleur ajustement au polynôme attendu :

$$y = y_0 + v_0 t - \frac{1}{2} g t^2$$

Utilisez la méthode aux moindres carrés pour obtenir les meilleures estimations des trois coefficients de cette loi et par la même, la meilleure estimation de g, à partir des cinq mesures du tableau :

X (temps t)	-2	-1	0	1	2
Y (hauteur h)	131	113	89	51	7

Hauteur (en cm) en fonction du temps (en dixième de seconde) d'un corps en chute. (*lm()*)

Exercice 20. Le rythme des désintégrations d'un échantillon radioactif décroît exponentiellement au cours du temps. Un étudiant suit cette décroissance exponentielle en enregistrant à l'aide d'un compteur le nombre de désintégrations durant le laps de temps de 15 s. Répétant cinq fois ses mesures à intervalle de 10 minutes, il obtient les résultats suivants :

X (temps t min)	10	20	30	40	50
Y (v désintégrations)	409	304	260	192	170

Nombre v(t) de désintégrations dans des intervalles de 15 s, en fonction du temps total écoulé.

Si la loi de désintégration est exponentielle, le nombre v(t) vérifie :

$v(t) = v_0 e^{-\lambda t}$
où λ et v_0 sont deux constantes inconnues.
Que vaut λ? Combien de désintégrations aurait-il relevé en 15 secondes à t=0 ? (*linéarisation du problème ou, à privilégier, package nls*)

Exercice 21. Un industriel s'intéresse à la tendreté des haricots qu'il mettra en boites. Pour

mesurer cette tendreté, on met les haricots en bottes (chaque botte ayant le même nombre de haricots) et on mesure le travail qu'il faut à une trancheuse pour couper la botte. L'indice de tendreté est proportionnel à ce travail. Les haricots peuvent provenir de fournisseurs différents F1, F2, F3 et F4. Les résultats sont les suivants :

F1	F2	F3	F4
8	9	10	6
9	10	9	8
9	9	12	8
6	10	11	7

On suppose que, pour tout $i \in \{1,..., 4\}$, l'indice de tendreté pour le fournisseur F_i est une variable X_i suivant la loi normale avec μ_i et σ_i inconnus.

1. Peut-on affirmer, au risque 5%, que la dispersion de l'indice de tendreté diffère selon les producteurs ?
2. Peut-on dire, au risque 5%, que l'indice de tendreté moyen diffère suivant la provenance des haricots ? (*bartlett.test(), anova(),pairwise.t.test() ou TukeyHSD()*)
3. Qu'aurions-nous du utiliser si l'hypothèse de normalité n'est pas vérifiée ?