



TD 6

Test statistique de comparaison de 2 distributions

Exercice 1: On considère 2 variables observées sur les mêmes individus.

T_i	45	49.5	52	54	61	63	74	80	83
P_i	8.98	8.14	6.67	6.08	5.90	5.83	4.68	4.20	3.72

Calculer un indice qui traduise le lien entre ces 2 variables et donner une interprétation de la valeur trouvée.

Exercice 2: Un fournisseur d'accès internet a modélisé par une loi de Poisson la distribution du nombre d'accès illicites à son serveur par minute. Il est admis que le paramètre de cette loi est 2.2. Les valeurs de cette loi de probabilité sont données dans le tableau suivant que l'on complètera :

Nombre d'accès illicites	0	1	2	3	4	5	> 5
Probabilité	0.1108	0.2438	0.2681	0.1967	0.1082	0.0476	

Un nouvel analyste pense que cette hypothèse de modélisation est erronée. Pour illustrer son propos, il note sur une période de 12 heures le nombre d'accès illicites au serveur pour chaque minute. Il obtient les résultats suivants :

Nombre d'accès illicites	0	1	2	3	4	5	6
Nb de minute	103	162	177	123	104	30	21

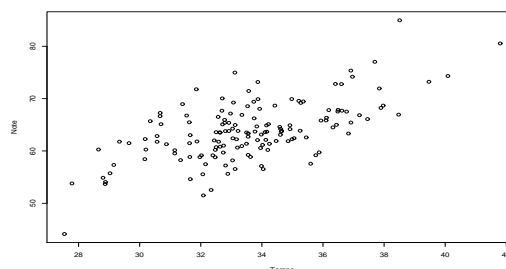
Dans un premier temps il calcule la moyenne (vous aussi!). Si l'hypothèse de loi de Poisson est vraie, à quoi peut servir cette valeur? Que pensez-vous de ce résultat?

Dans un second temps il remet en cause l'hypothèse de la loi. Proposez une démarche statistique complète qui permette de justifier cette idée (le type de test, les hypothèses, la statistique de test, la zone de rejet). On prendra une erreur de 1^{re} espèce de 2.5%.

Menez à bien l'ensemble des calculs et proposez une conclusion.

Proposez une nouvelle erreur de 1^{re} espèce qui ne change pas la conclusion mais qui améliore la qualité du test (précisez en quel terme).

Exercice 3: On a mesuré sur 150 patients le temps mis pour exécuter une tâche particulière ainsi que la note (sur 100) à un test de coordination. On trace le graphe de ces 2 variables :



On cherche alors à identifier l'existence d'une liaison. On a noté : $\bar{t} = 33.5$, $\hat{\sigma}_t = 2.338$, $\overline{no} = 63.5$, $\hat{\sigma}_{no} = 5.56$, $\bar{t} * \overline{no} = 2136.2$. Proposer un indice et une démarche. Conclure.

Exercice 4:

1. On cherche à déterminer si les étudiants ont le même taux de redoublement dans leur cursus secondaire selon la filière choisie. On note alors les résultats de l'enquête du mois de septembre dans le tableau suivant :

Filière	Sciences	Droit	Psycho	Médecine
"Redoublant" : NON	49	52	63	61
"Redoublant" : OUI	131	128	149	151

Peut-on considérer que la filière et le caractère étudié sont liés?

2. Une autre question est de déterminer s'il y a une différence entre le score évalué sur les étudiants Garçons et les étudiants Filles lors de l'enquête du mois de septembre.
- Le score moyen pour les garçons (ils sont 380) est : $\overline{s^G} = 328$ et leur dispersion $\sum_{i=1}^{380} (s_i^G - \overline{s^G})^2 = 492000$.
 - Le score moyen pour les filles (elles sont 404) est : $\overline{s^F} = 343.5$ et leur dispersion $\sum_{i=1}^{404} (s_i^F - \overline{s^F})^2 = 524000$.

Peut-on considérer que le score de ces 2 groupes est différent?

Exercice 5: (reprendre l'énoncé de l'exercice 5 TD1)

1. Pour 310 clients qui sont passés aux caisses automatiques, on a noté leur heure de passage. On a obtenu les réponses suivantes :

Heure	entre 13 et 14 h	entre 14 et 15 h	entre 15 et 16 h	entre 16 et 17 h
Effectifs	94	67	59	90

Peut-on considérer que la répartition des passages aux caisses automatiques est uniforme durant les 4 créneaux horaires de l'après-midi?

2. Aux 310 clients "Moins de 10" de l'après-midi, on a comparé leurs dépenses à 400 clients "Moins de 10" passés en caisse après 17h.

On a obtenu pour :

- Les 310 clients "Moins de 10" de l'après-midi, une dépense moyenne $\overline{d^a} = 11.5$ euros et une dispersion des dépenses $\sum_{i=1}^{310} (d_i^a - \overline{d^a})^2 = 8300$.
- Les 400 clients "Moins de 10" d'après 17h, une dépense moyenne $\overline{d^b} = 13.2$ euros et une dispersion des dépenses $\sum_{i=1}^{400} (d_i^b - \overline{d^b})^2 = 10800$.

Peut-on considérer que la dépense moyenne de ces 2 groupes de clients est différente?

Exercice 6: Dans une séquence ADN de longueur 3500, formée des lettres A, C, G, T, on compte le nombre d'occurrences de chacune des lettres. Et on cherche à éprouver l'hypothèse que chaque lettre apparaît de façon équiprobable. Construire le test approprié en spécifiant clairement les hypothèses, la statistique du test et la zone de rejet.

Exercice 7: On a mesuré une variable continue X sur 4 groupes disjoints d'individus, et on a obtenu les résultats suivants :

	Taille du groupe	Moyenne	Variance
Groupe 1	380	46.4	300.9
Groupe 2	280	45.6	298.4
Groupe 3	455	52.2	294.0
Groupe 4	165	47.1	320.8

On se pose naturellement la question de savoir si ces 4 groupes ont des résultats similaires au vu de cette variable X . Construire le test approprié en spécifiant clairement les hypothèses, la statistique du test et la zone de rejet. Proposer votre conclusion.