Partie 2 : statistiques analytiques

Cette partie, contrairement aux statistiques descriptives, ne se contente pas de la simple description des observations, mais aide à tester des hypothèses, à concevoir des constatations et de prendre des décisions.

Deux hypothèses sont posées soit H0 et H1, le test statistique va être l'outil qui permet de trancher ; une seule qui est vraie.

H0 est appelée l'hypothèse nulle, H1 est l'hypothèse alternative La décision consiste à retenir H0 ou H1.

Cette partie sera abordée comme suit

A. Comparaison de deux moyennes
B. Analyse de variance à un facteur
C. Analyse de variance à facteurs multiples

A. **Comparaison de deux moyennes**

A.1. Tests paramétriques (Test de Student)

a. Comparaison d'une moyenne à une valeur référence ou moyenne théorique

b. Comparaison de moyenne sur deux échantillons appariés

c. Comparaison de moyenne de deux échantillons indépendants

A.2. Tests non paramétriques (Test de Chi 2)

a. Comparaison de moyennes observes sur échantillons indépendants

b. Comparaison de moyenne sur échantillons appariés

A.1. Test de Student

Conditions d'application:

Le caractère de l'échantillon étant supposé aléatoire, l'hypothèse de normalité de la variable doit être vérifiée.

Le test de Student est utilisé dans trois cas différents :

a. Comparaison d'une moyenne à une valeur référence

b. Comparaison de moyennes sur deux échantillons appariés (liés)

c. Comparaison de moyennes de deux échantillons indépendants

Quelque sois le cas, la réalisation suit les mêmes étapes :

- L'hypothèse nulle H0 : démarrer avec l'hypothèse qu'il n'ya pas de différences significatives entre les deux moyennes comparées

  L'hypothèse alternative :H1 : il y a une différence significative entre les deux échantillons

- Calcul du $t_{.cal}$ calculé selon une formule spécifique à chaque cas

- Calcul de $t_{.tab}$ (t de table) : cette valeur est lue dans une table de Student a l'aide de deux valeur(α=0.05, ddl=(voir le ddl dans chaque cas))
- Comparaison entre $t_{.cal}$ et $t_{.tab}$
- Conclusion (accepter H0 et rejeter H1 ou le contraire)

***Cas n°1** : Comparaison d'une moyenne à une valeur référence ou moyenne théorique*

Il s'agit de comparer une **moyenne d'un échantillon** à une **moyenne théorique** (µ).Le but est de vérifier si notre échantillon provient bien d'une population avec la moyenne théorique, μ, ou non

Calcul :

Soit X une variable aléatoire distribuée selon une loi normale, la variable aléatoire définie ci-dessus suit une loi de Student avec n - 1 degrés de liberté.

$$\textbf{t cal} = \frac{\overline{X} - \mu}{\sqrt{\frac{S^2}{n}}}$$

- $\overline{x}$: la moyenne de l'échantillon
- µ : la moyenne théorique ou la valeur spécifiée
- S^2 : la variance de l'échantillon
- n : est la taille de l'échantillon

la valeur t tab correspondant au **risque alpha** = 5% pour un degré de liberté : ddl= n−1

On compare la valeur calculée de t (cal) avec la valeur critique appropriée de t-table avec n - 1 degrés de liberté.

- Si t cal est inferieur à t tab : H0 est acceptée

-Si t cal est supérieure à t tab : H0 est rejetée et on accepte H1.

***Cas n° 2 :** Comparaison de moyenne sur deux échantillons appariés*

Le test de Student sert aussi à comparer les moyennes de deux populations, dont chaque élément de l'une est en relation avec un élément de l'autre.

Par exemple 30 patients ont reçu un traitement pendant1 mois. On se pose la question à savoir si le traitement a un impact sur le poids des patients ou pas. Le poids des sont donc été mesuré avant et après traitement. Ce qui nous donne 30 séries de valeurs avant traitement et 30 après traitement provenant des mêmes individus.

Il s'agit bien dans cet exemple, d'un **test de Student apparié** car les deux séries de valeurs ont un lien (les souris). Pour chaque souris, on a deux mesures (l'une avant et l'autre après traitement).

Calcul :

Soit *xij* l'observation *j* pour la paire *i* (j = 1,2 et i = 1,2,...,n). Pour chaque paire d'observations on calcule la différence di = xi2 - xi1. Le test statistique est défini par :

tcal = $\frac{\bar{d}}{\sqrt{\frac{S_d^2}{n}}}$

- *n est le nombre de paires d'observations*
- *d est la moyenne des différences entre les observations et*
- *Sd² la variance des différences.*

d.d.l=n−1

***Cas n°3 :** Comparaison de moyennes de deux échantillons indépendants*

Etant donné deux échantillons de taille n1 et n2. Dans ce cas les deux échantillons sont indépendants entre eux.

Calcul :

On calcule la valeur t cal comme suit

T cal = $\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}}$

- *x1 et x2 sont les moyennes des deux échantillons*
- *Sp² la variance commune.*

- *n 1* : est la taille de l'échantillon 1
- *n 2* : est la taille de l'échantillon 2

La valeur t tab correspondant au **risque alpha** = 5% pour un degré de liberté : ddl= n1+n2−2

*Si la valeur de t calculé (|t|) est supérieure à la valeur de t de table, alors la différence est significative. Dans le cas contraire, elle, ne l'est pas. Le degré de significativité (ou p-value) correspond au risque indiqué par la table de Student pour la valeur |t|.

Table de Student t

ν	α					
	0.100	0.050	0.025	0.010	0.005	0.001
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.611
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
100	1.290	1.660	1.984	2.365	2.626	3.174
∞	1.282	1.645	1.960	2.326	2.576	3.090

La table de Student donne les valeurs $t_{(\alpha,\nu)}$ telles que

$$P\{T > t_{(\alpha,\nu)}\} = \alpha$$

A .2. Le Test non paramétrique Chi 2 ou Khi 2

Le test d'indépendance du Chi 2 sert à apprécier l'existence ou non d'une relation entre deux caractères au sein d'une population, lorsque ces caractères sont qualitatifs où lorsqu'un caractère est quantitatif et l'autre qualitatif, ou bien encore lorsque les deux caractères sont quantitatifs mais que les valeurs ont été regroupées.

Pour les caractères qualitatifs la notion de moyenne n 'existe pas donc il faux prendre e compte la distribution dans son ensemble

Par **exemple** si on veut vérifier si la répartition des groupes sanguins dépend du sexe

Les étapes à suivre sont comme suit :

1- On pose l'hypothèse H0 : "Il n'y a pas de relation entre la répartition des groupes sanguins et le sexe

2- Calcul :

On détermine la valeur du chi2 calculé (chi2 $_{cal}$)comme suit

chi2 $_{cal}$= Σ (effectifs observes-effectifs théoriques)^2/ effectifs théoriques

3- On détermine le chi2 table, Cette valeur est lue dans une table du test du Chi-2 a l'aide de deux valeur(α=0.05, ddl=n-1)

4- comparer chi2 $_{cal}$ et chi 2 $_{tab}$

-si chi2 $_{cal}$ inferieur à chi 2 $_{tab}$: y'a une indépendance entre la répartition des groupes sanguins et le sexe

-si chi2 $_{cal}$ supérieur à chi 2 $_{tab}$: y'a une dépendance entre la répartition des groupes sanguins et le sexe

Avec un risque d'erreur de α=5%.

5- explication de la signification

Ce test permet de déterminer la dépendance mais ne donne pas la puissance de cette dépendance.

- Le Chi 2 aussi peut aussi être un test d’ajustement pour comparer la distribution de fréquences d’un échantillon par rapport à la distribution de fréquences attendu c.a.d évaluer à quel point les données que vous obtenez lors d’une expérience sont bien ajusté aux données théoriques attendus.

Table du chi-carré χ^2

ν	α						
	0.900	0.700	0.500	0.300	0.100	0.050	0.010
1	0.016	0.15	0.46	1.07	2.71	3.84	6.63
2	0.21	0.71	1.39	2.41	4.60	5.99	9.21
3	0.58	1.42	2.37	3.67	6.25	7.81	11.34
4	1.06	2.19	3.36	4.88	7.78	9.49	13.28
5	1.61	3.00	4.35	6.06	9.24	11.07	15.09
6	2.20	3.83	5.35	7.23	10.65	12.59	16.81
7	2.83	4.67	6.35	8.38	12.02	14.07	18.48
8	3.49	5.53	7.34	9.52	13.36	15.51	20.09
9	4.17	6.39	8.34	10.66	14.68	16.92	21.67
10	4.87	7.27	9.34	11.78	15.99	18.31	23.21
11	5.58	8.15	10.34	12.90	17.28	19.68	24.73
12	6.30	9.03	11.34	14.01	18.55	21.03	26.22
13	7.04	9.93	12.34	15.12	19.81	22.36	27.69
14	7.79	10.82	13.34	16.22	21.06	23.69	29.14
15	8.55	11.72	14.34	17.32	22.31	25.00	30.58
16	9.31	12.62	15.34	18.42	23.54	26.30	32.00
17	10.09	13.53	16.34	19.51	24.77	27.59	33.41
18	10.87	14.44	17.34	20.60	25.99	28.87	34.81
19	11.65	15.35	18.34	21.69	27.20	30.14	36.19
20	12.44	16.27	19.34	22.78	28.41	31.41	37.57
25	16.47	20.87	24.34	28.17	34.38	37.65	44.31
30	20.60	25.51	29.34	33.53	40.26	43.77	50.89
35	24.80	30.18	34.34	38.86	46.06	49.80	57.34
45	33.35	39.58	44.34	49.45	57.50	61.66	69.96
55	42.06	49.06	54.33	59.98	68.80	73.31	82.29
65	50.88	58.57	64.33	70.46	79.98	84.82	94.42
75	59.79	68.13	74.33	80.91	91.06	96.22	106.39
85	68.77	77.71	84.33	91.32	102.08	107.52	118.24
95	77.82	87.32	94.33	101.72	113.04	118.75	129.97
120	100.62	111.42	119.33	127.61	140.23	146.57	158.95

La table du chi-carré donne les valeurs $\chi^2_{(\alpha,\nu)}$ telles que

$$P\{\chi^2 > \chi^2_{(\alpha,\nu)}\} = \alpha$$