Analyse des données

I. Introduction.

- Analyse univariée : C'est l'étude de variables une à une (tendance moyenne centrale : Mode, médiane et dispersion : Variance, écart-type).
- Analyse bivariée : Etude des relations entre deux variables. Calculs d'indices (Khi², n, etc…).
- Analyse multivariée : Analyse de plus de deux variables simultanément.

II. Décrire les variables : Le tris à plat.

- Profil des réponses (description de la population, quel écart par rapport à la population mère ?).
- Distribution des réponses (% par rapport à l'ensemble des sujets des réponses).
- Y a t il des non réponses ? Si oui pourquoi ? (question défectueuse, indiscrète, tendancieuse ou oubliée tout simplement).

III. Les mesures de tendance centrale.

Il faut résumer l'ensemble des données : Mode, médiane, moyenne.

- Mode (Mo) : La ou les valeurs, modalité(s) la plus fréquente. Les distributions univariées, bivariées ou multivariées sont modales.
- Médiane (Med) : On divise une série statistique ordonnée en deux groupes comprenant chacun 50% des données.
- La moyenne (M) : C'est la somme des valeurs observées divisées par N.

Pour déterminer la médiane d'une variable qualitative ordinale :

- On classe les données par ordre croissant.
- Pour une série de n observations, la médiane correspond à la modalité qui occupe le rang n/2 (si le nombre d'observations est pair) ou (n+1)/2 si le nombre d'observations est impair.

Par exemple : On a 25 sujets, on leur propose une tâche de motricité, on code la rapidité d'éxécution : Très lent (TL) ; Lent (L), Rapide (R) ou Très Rapide (TR). On commence par classer les données par ordre croissant :

TL TL TL TL TL TL L L L L L L L | L R R R R R R TR TR TR TR : Médiane = (n+1)/2 soit 26/2 = 13

IV. La dispersion.

1. Calculer la moyenne.
2. Soustraire la moyenne de chaque score.
3. Elever au carré la différence obtenue pour chaque score.
4. Additionner ces différences au carré (On obtient la Somme des Carrés : SC).
5. Diviser SC par le nombre de cas : On obtient la variance.
6. Extraire la racine carrée de la variance : On obtient la déviation standard ou écart-type.

V. Forme de la distribution.

Illustration tirée du logiciel « Sassi. » utilisé par le prof.
Coefficient de Kurtosis (G2 de Fisher) :

Valeur du G2	Interprétation
0	Courbe normale
Proche de 0	Quasi-normale
>>> 0	Pointue
<<<0	Aplatie
Proche de -1.3	Uniforme plate
<< -1.3	En U.

Formule :

$$\left[\frac{\sum (\text{valeur-moyenne})^4}{\text{Effectif}} \right] -3$$

Le coefficient d'asymétrie (Skewness) de Fisher (G1) :

Valeur du G1	Interprétation
G1<0	Plus de dispersion pour les valeurs < à la moyenne que pour les valeurs >.
G1 =0	Les valeurs inférieures et supérieures à la moyenne distribuent de la même manière (mode, médiane, moyenne se superposent).
G1 > 0	Plus de dispersion pour les valeurs > à la moyenne que pour les valeurs <.

Formule

$$\left[\frac{\sum (\text{valeur-moyenne})^3}{\text{Effectif}} \right]$$

Les formules sont un peu plus compliquées que ça, je vous conseille d'aller les trouver dans des cours de statistiques, je ne suis pas arrivée à les recopier en entier.

VI. Statistiques inférentielles.

- Aller plus loin que la description en se basant sur les lois probabilistes.
- Utilise un degré de significativité ($p < .05$; $p < .01$ etc…).

Tests paramétriques.		Sujets indépendants	Mesures répétées
		T de Student pour El	T de Student pour échantillon par paires.
Tests non paramétriques.	Khi² d'ajustement	Khi² d'association Mann Whitney	Test de Wilson. Test du signe.

Coefficient De corrélation	Nominal	Ordinal	Intervalle
	Phi, V de Cramer.	Rhô de Spearman. Tan de Kendall	R de Peason

A. Tester l'indépendance de mesures nominales : Le khi².

1. Le tableau de contingence.

On étudie le lien entre le style parental sur le jeu auprès d'un groupe de 30 enfants âgés de 7 à 8 ans.

1. Un questionnaire a été proposé aux parents afin d'évaluer leurs style éducatifs et d'affecter les enfants par la suite dans trois catégories.
2. On observe ensuite les enfants jouer et on évalue leur style de jeu dominant : Coopératif ou compétitif. Les réponses sont consignées dans un tableau de contingence.

	Coopératif	Compétitif	T
Permissif	7	15	22
Intermédiaire	21	9	30
Autoritaire	4	28	32
T	32	52	N =84

2. Tableau d'indépendance.

On calcule les effectifs théoriques :

	Coopératif	Compétitif	T
Permissif	(33*22)/84 = 83	(22*52)/84 = 13.6	22
Intermédiaire	(30*32)/84 = 11.4	(30*52)/84 =18.5	30
Autoritaire	(32*32)/84 =12.1	(32*52)/84 = 19.8	32
T	32	52	N =84

Les chiffres correspondant aux effectifs théoriques sont le résultat de la multiplication de l'effectif total par ligne * l'effectif par colonne divisé par l'effectif total.

3. Tableau d'écarts à l'indépendance.

On fait effectif observé – effectif théorique (calculé à partir dans le tableau d'indépendance).

	Coopératif	Compétitif	T
Permissif	7-8.3 = -1.3	15-13.6 = 1.4	22
Intermédiaire	24-11.4 = 9.6	9-18.5 = -9.5	30
Autoritaire	4-12.1 = -8.1	8.2-19.8 = 8.2	32
T	32	52	N =84

On constate que la somme des écarts à la moyenne est bien égale à zéro (c'est toujours le cas et ça doit toujours l'être).

4. Tableau des carrés des écarts à l'indépendance.

On reprend les valeurs obtenues dans le tableau des écarts à l'indépendance, on les élève au carré et on les divise par les effectifs théoriques :

	Coopératif	Compétitif	T
Permissif	(7-8.3)²/8.3 = 0.20	(15-13.6)²/13.6 = 0.14	0.34
Intermédiaire	(24-11.4)²/11.4= 8.08	(9-18.5)²/18.5 = 4.87	12.95
Autoritaire	(4-12.1)²/12.1 = 5.42	(8.2-19.8)²/19.8 = 3.39	8.81
T	13.7	8.4	Khi² calculé = 22.1 (somme des lignes ou des colonnes)

Si le Khi² calculé (à l'aide du tableau précédent) est supérieur au Khi² lu (dans la table), l'effet est significatif.
On lit la valeur du Khi² en fonction du degré de liberté et du seuil de significativité choisi.
Le Ddl (degré de liberté) se calcule comme suit : (c-1)*(l-1) soit (nombre de colonnes -1) * (nombre de lignes - 1).
En ce qui concerne le degré de significativité on choisi souvent 5% soit 0.05 (7ème colonne en partant de la gauche dans la table).
Dans cet exemple, le degré de liberté est de 2, on a choisi un seuil de 5%, on lit la valeur correspond à la ligne 2 et à la colonne 0.05 et on constate qu'elle est inférieure à celle obtenue par les calculs.
Au seuil de 5%, avec un degré de liberté de 2, on observe un effectif significatif.

Conclusion.

On observe un lien significatif entre le style éducatif parental et le type de jeu des enfants, le résultat a plus de 99.9% de significativité, l'hypothèse a une chance d'être erronée dans 0.01% des cas.

5. Condition d'application du Khi².

Pour un tableau à 4 cases :

- Il faut un effectif théorique supérieur à 9.
- On applique une correction de Yates (on soustrait 0.5 avant d'élever au carré les écarts à l'indépendance), qui fait chuter la valeur du Khi².

Pou un tableau à plus de 4 cases il n'y a pas d'effectif théorique inférieur à 1 et moins de 20% de cases où l'effectif théorique est inférieur à 5.

B. Les corrélations : L'exemple du r de Bravais-Pearson.

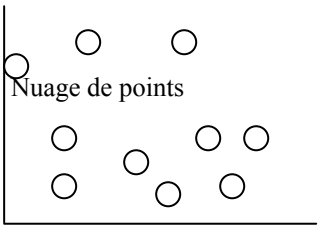
1. La force du lien.

Valeur absolue de corrélation	Relation linéaire entre les deux valeurs
Entre 0 et .20	Nulle à faible
De 0.20 à .40	Faible à modérée
De .40 à .70	Modérée à élevée
De.70 à .90	Elevée à très élevée
De .90 à 1	Très élevée à parfaite (correspond aux valeurs qu'on obtient lorsqu'on corrèle un item avec lui-même).

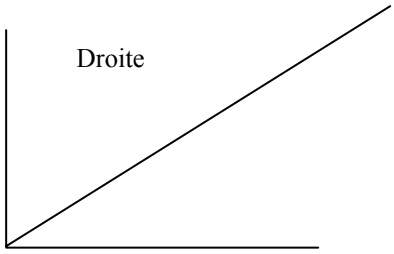
NB = r est différent du pourcentage de variance expliquée (r2), par exemple si on trouve r = .40, le pourcentage de variance expliquée correspond à r² soit 16%

2. Le sens du lien : Positif/Négatif.

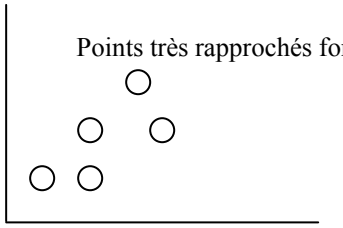
a. Corrélation nulle, r = 0



b. Corrélation parfaite, r = 1



c. Corrélation très élevée, r = 0.9



d. Corrélation élevée négative, r = .70

Points plus espacés que schéma précédent et sens inversé
Les plus hauts sont au début et les plus bas au bout de l'axe.

e. Corrélation parfaite négative, r= .1

C'est une droite inversée.

f. Corrélation moyenne, r = 0.6

Les points sont relativement éparpillés.

3. La significativité du lien.

Par exemple : Une matrice d'inter corrélations.

	A	B	C	D	E	F
A	\	.20	.14	.49***	.15	.12
B	\	\	.29*	52***	.35**	.33*
C	\	\	\	.29*	31*	34**
D	\	\	\	\	.29*	.36**
E	\	\	\	\	\	.51***
F	\	\	\.	\	\	\

On remarque que l'item A est mal corrélé donc on le supprime.
On marque les corrélations moyennes (bonnes) avec 3 * (***).
On marque les corrélations moins bonnes avec 2* (**)
On marque les corrélations faibles avec une étoile (*)
On marque les corrélations quasi nulles avec rien.

4. Calculer un r.

Formule :

$$\frac{\sum (\text{valeur de x} - \text{moyenne de x}) - (\text{valeur de y} - \text{moyenne de y})}{\sqrt{\sum (\text{valeur de x} - \text{moyenne de x})^2 * \sum (\text{valeur de y} - \text{moyenne de y})^2}}$$

La significativité du coefficient de corrélation.

- Objectif : Généraliser à l'ensemble de la population.
- Procédure :
  - Déterminer le Ddl (N-2).
  - Lire le seuil de risque correspondant.

Si r calculé est supérieur à r lu l'effet est significatif.

5. Avec Excel :

A	B
1	3
3	4
4	5
5	4
6	3
4	3
7	3
7	4
1	5
2	2

On insère une fonction =PEARSON(A1:A10;B1:B10)

C. Tester l'homogénéité d'une échelle : L'alpha de Cronbach.

Voir polycop' donné par le prof en TD.

D. Comparer les moyennes : Le test t.

Le T test pour échantillons indépendants (la mesure prise sur les unités d'un groupe ne doit pas influer sur le résultat de la mesure prise sur les unité de l'autre groupe).

1. Objectif.

Evaluer si la différence observée entre deux moyennes résulte des fluctuations dues au hasard ou si un autre facteur peut l'expliquer.

2. Procédure.

On utilisera un modèle probabiliste appelé « distribution t », et on comparera la probabilité de l'événement observé avec celle prédite par le modèle probabiliste en supposant que seul joue le hasard.
Exemple : La pratique judiciaire professionnelle et la punitivité pénale.

Hypothèse : Les sujets issus du milieu judiciaire (Groupe J) sont moins punitifs que les sujets n'étant pas acclimatés à ce milieu (Groupe PG).

Utilisation d'une échelle de punitivité pénale en 10 items (construction et test préalables).

On suppose d'abord qu'il n'y a pas de différence (c'est l'hypothèse nulle : les moyennes des groupes sont les mêmes) et on tente de déterminer si les données sont compatibles avec cette hypothèse en comparant la moyenne de chacun des échantillons.

Si la moyenne du Groupe J est nettement différente de celle du groupe PG, on affirmera que la pratique sociale spécialisée est associée au jugement pénal.

Avant de généraliser cette conclusion à l'ensemble de la population, on effectuera un test d'hypothèse sur la différence entre les deux moyennes.

3. Etapes de la recherche.

a. La formulation des hypothèses.

Dans un test de la différence entre deux moyennes, les hypothèses peuvent être formulées de trois façons selon qu'on veut que montrer que la moyenne du premier groupe (J) est différente, inférieure ou supérieure à la moyenne du second groupe (PG).

On supposera ici l'existence d'une différence : Le groupe J devrait avoir un score de punitivité pénale inférieur au groupe PG.
Nous procéderons à un test bilatéral.

b. Le choix du seuil de significativité.

Au seuil de 5% soit 0.05.

c. Vérification des conditions d'application.

- Le test que nous faisons s'applique lorsque les échantillons sont de petite taille (inférieur à 30).
- Les échantillons doivent avoir été prélevés au hasard et de manière indépendante au sein d'une population pour laquelle la caractéristique observée dans chacun des deux groupes correspond au modèle de la loi normale et pour laquelle les écarts types des groupes sont homogènes.

d. Le calcul de la V d'écart.

La V d'écart (t) mesure la différence entre les deux moyennes pondérées par un écart type ajusté de façon à tenir compte des écart type de la taille des échantillons.
Variance : Moyenne des carrés des écarts-types.
Sd = Racine carré de la variance.

e. Formule du t :

$$T = \frac{\text{Moyenne échantillon} - \text{Moyenne population}}{\sqrt{\frac{(n1-1)\,s^2 1 + (n2-1)\,S^2 2}{N1+n2-2}} * \sqrt{1 + \frac{1}{n1}\ \frac{1}{n2}}}$$

f. La détermination de la valeur critique.

On trouve la valeur critique dans une table de valeurs critiques. Cette valeur dépend :
- Du seuil de significativité alpha.
- Du nombre de degrés de liberté.
- De la nature du test (bi ou unilatéral).

La courbe de la loi de Student, d'où sont tirées les valeurs critiques, ressemble à la courbe de la loi normale.

g. La formulation de la règle de décision.

La règle de décision, dans un test de la différence entre deux moyennes, est fonction de l'hypothèse alternative retenue.
On rejette l'hypothèse nulle si elle paraît statistiquement incompatibles avec les résultats.

h. La décision.

Au seuil de signification fixe et en vertu de la règle de décision, on décide de rejeter ou de maintenir l'hypothèse nulle, on acceptera l'hypothèse alternative et on dira que le résultat est significatif au seuil de alpha.

Retour à l'exemple :

	J	PG
N	8	8
Moyenne	20.1	17.4
Ecart-type	2.4	1.6

Le nombre de Ddl est v = nl + nc – 2 soit 8 +8 -2 = 14.

$$T = \frac{20.1-17.4}{\sqrt{\frac{(8-1)\,2.4^2 + (8-1)\,1.6^2}{8+8-2}} \sqrt{\frac{1}{8} + \frac{1}{8}}}$$

Pour v = 14 et alpha = 5%, nous obtenons t critique = 1.76 puisqu'il s'agit d'un test bilatéral. Nous constatons que t = 2.65>1.16 = tc.
On rejette donc l'hypothèse nulle, le groupe J est plus clément que l'hypothèse que le groupe PG.