Statistiques descriptives : Résumés et exercices

Jean-Marc Meunier

Institut d'Enseignement à Distance de
l'Université de Paris 8.

DEUG de psychologie première année

STATISTIQUES DESCRIPTIVES

Résumés et exercices

Jean-Marc Meunier

Référence : R 2440 T
Classe 321

INTRODUCTION.

Cette introduction est surtout une mise en garde contre la tentation de croire que l'étude de ce document puisse permettre de se dispenser de l'étude du cours proprement dit. Le propos de ce document est de vous proposer une aide à l'étude du cours. Il est organisé dans le respect de la structure de votre cours. Vous y trouverez :

- Une définition simple des principaux concepts.
- Un résumé du cours
- Quelques exercices.
- Les principaux pièges à éviter.
- Une foire aux questions.

. La réalisation des exercices proposés n'a aucun caractère obligatoire, mais est vivement conseillée surtout dans les parties du cours que vous avez du mal à appréhender. Ces exercices ne doivent pas être envoyé à la correction. Pour chacun d'eux, vous trouverez un corrigé vous permettant de vous évaluer.

Avant de commencer, voici succinctement quelques rappels d'algèbre, pour ceux qui en auraient besoin :

Rappels d'algèbre.

Règles de priorité des opérations.

Dans tous les cas, on commence par l'intérieur des parenthèses. On doit réaliser les opérations en commençant par les élévations à la puissance, puis les multiplications et les divisions, puis les additions et les soustractions.

Addition de nombres réels :

- Lorsque les deux nombres sont de même signe, on fait la somme des deux nombres, le résultat est de même signe que ces deux nombres **:**

(-2)+(-2)=-4 et (+2)+(+2)=+4

Lorsque les deux nombres sont de signe**s** différent**s**, on fait la différence des deux nombres et le résultat est de même signe que le plus grand des deux nombres.

(-2)+(+5)=3 et (+2)+(-5)=-3

Produit de nombres réels:

Lorsque les deux nombres sont de même signe, le résultat est de signe positif (de ce fait un carré n'est jamais négatif).

Lorsque les deux nombres sont de signes différents, le résultat est de signe négatif.

Addition de fractions.

Pour additionner des fractions, il faut réduire au même dénominateur et ensuite additionner les numérateurs.

$$\frac{1}{2}+\frac{1}{3}=\frac{(1*3)}{(2*3)}+\frac{(1*2)}{(3*2)}=\frac{3}{6}+\frac{2}{6}=\frac{5}{6}$$

Produit de fractions

On multiplie en ligne les numérateurs entre eux et les dénominateurs entre eux.

$$\frac{1}{2} * \frac{1}{3} = \frac{1*1}{2*3} = \frac{1}{6}$$

Division par une fraction

Pour diviser par une fraction, on multiplie pas l'inverse.

$$2 : \frac{1}{3} = 2 * \frac{3}{1} = 6$$

Identités remarquables

$$(a+b)^2 = a^2 + b^2 + 2ab$$

$$(a-b)^2 = a^2 + b^2 - 2ab$$

$$(a+b)(a-b) = a^2 - b^2$$

Quelques mots sur les écritures mathématiques.

Les individus statistiques sont notés i.

Les modalités de la variable sont notés u.

Les observations sont notées x_i, où i désigne la ligne correspondant à un individu statistique.

Le total des observations est noté T ou Σx_i.

Les effectifs d'une modalité sont notés n_u, où u désigne une modalité particulière.

L'effectif total (nombre d'observation) est noté N.

La fréquence d'une modalité est notée f_u, où u désigne une modalité particulière.

Le symbole Σ (Somme) signifie qu'on doit appliquer la formule qui le suit à toutes les lignes du tableau, puis faire la somme des résultats de chaque ligne.

$\sum x_i$ veut dire qu'il faut faire la somme des valeurs de x

$\sum x_i^2$ veut dire qu'il faut élever chaque observation au carré, puis faire la somme de ces carrés.

LES NOTIONS DE BASE.

RÉSUMÉ

Les méthodes statistiques portent sur des ensembles d'individus statistiques. L'ensemble d'individus étudiés est appelé "échantillon". Cet échantillon est un sous-ensemble d'un ensemble plus large appelé "population".
On étudie sur ces échantillons certaines dimensions appelées "variables". Les différentes valeurs que peuvent prendre ces variables sont appelées modalités. Une variable a forcément plus d'une modalité.
Une variable est caractérisée par son échelle de mesure (nominale, ordinale, d'intervalle ou numérique) et éventuellement son statut (variable dépendante vs[1] variable indépendante). Les variables indépendantes sont également appelées "facteurs".

- ✓ Une variable nominale est caractérisée par le fait que ses modalités n'entretiennent pas de relation d'ordre.
- ✓ Une variable ordinale est caractérisée par des modalités ordonnées entre elles.
- ✓ Une variable numérique a des modalités ordonnées et un intervalle constant entre modalités.

Une observation est la mise en relation d'une modalité d'une variable avec un individu statistique. L'ensemble des observations est appelé "protocole".
Un protocole est caractérisé par l'éventuelle présence d'une structure. Il sera dit structuré si on peut mettre en évidence une relation d'emboîtement ou de croisement entre les individus statistiques et les facteurs.
Les objectifs de la méthode statistique sont :

- Résumer un protocole ou une distribution.
- Situer un sujet dans une distribution.
- Comparer des groupes d'observations.
- Évaluer l'existence d'une relation.

LES PIÈGES À ÉVITER

Confondre les individus statistiques et les modalités de la variable. Il arrive parfois qu'une lecture un peu trop rapide de l'énoncé conduise les étudiants à confondre ces deux notions.

[1] note : vs signifie Versus, c'est-à-dire « par opposition à ».

Elles sont pourtant bien distinctes, puisque les modalités de la variable caractérisent les individus statistiques et non l'inverse.

Confondre les modalités d'une variable avec les variables. Les variables sont des dimensions qui servent à caractériser les individus statistiques, les modalités sont les valeurs que peuvent prendre ces dimensions. L'erreur qui consiste à les confondre va souvent de pair avec une autre erreur qui consiste à déclarer comme variable une caractéristique constante chez les sujets. Pour l'éviter, il faut toujours se demander, lorsqu'on croit avoir identifié une variable, quelles sont les différentes valeurs que cette variable peut prendre. Si on ne peut pas répondre à cette question en identifiant plus d'une valeur, c'est qu'on est en train de se tromper.

Déclarer comme variable numérique une variable nominale dont les modalités sont représentées par des chiffres. Pour des commodités de codage, il arrive souvent qu'on représente les différentes modalités d'une variable nominale par des chiffres. C'est le cas par exemple lorsqu'on fait passer un questionnaire et que les réponses sont codées par leur numéro. Pour éviter ce genre d'erreur, il faut toujours se demander ce que représentent les chiffres. Ce ne sont parfois que de simples étiquettes utilisées parce qu'elle tiennent moins de place dans un tableau que des mots ou des phrases.

Déclarer des variables dépendantes ou indépendantes dans un protocole univarié. La question du statut des variables n'a de sens que si on a au moins deux variables.

Ne déclarer que des variables indépendantes ou que des variables dépendantes dans un protocole multivarié. Le statut d'une variable n'a de sens que par opposition à l'autre statut. On ne peut en effet avoir seulement des variables indépendantes (elles sont censées influencer quelle variable ?) ou seulement des variables dépendantes (elles sont censées être influencées par quelle variable?). Si, dans un énoncé, vous trouvez une variable indépendante, alors il existe forcément quelque part une variable dépendante, et inversement.

Déclarer une structure de protocole avec des facteurs qui n'ont pas été identifiés au préalable. Pour un enseignant, cette capacité de l'étudiant est surprenante. Elle signale un défaut du contrôle de la cohérence de son propos. C'est le genre de chose qui passe facilement à la trappe sous la pression de l'examen. Elle consiste par exemple à déclarer une structure d'emboîtement, alors que le facteur emboîtant n'est pas identifié en tant que tel. La description d'un protocole doit être cohérent dans son ensemble. En gardant cela à l'esprit, on évite facilement ce genre d'erreur.

S'ENTRAINER

Pour chacun des comptes-rendus de recherche suivants, nous vous demandons de décrire le protocole (individus statistiques, variables et structure du protocole) et les objectifs statistiques de la recherche.

Exercice 1.

Les performances en lecture de soixante-seize enfants bénéficiant de deux méthodologies didactiques contrastées (une approche idéo-visuelle[2] pure et une approche partiellement phonique[3]) sont comparées au terme d'une étude longitudinale de vingt-huit mois (de la fin de grande section de maternelle au début du cours élémentaire 2ème année. Les élèves bénéficiant d'une didactique[4] idéo-visuelle obtiennent des scores nettement inférieurs à ceux des autres élèves lors des évaluations nationales de CE2 malgré des performances initiales équivalentes en fin de scolarité maternelle. Leurs vitesses d'identification des mots écrits sont plus lentes que celles des élèves bénéficiant d'une didactique phonique rénovée. L'absence d'enseignement du code grapho-phonologique[5] apparaît comme un obstacle à l'apprentissage de la lecture au cycle 2 et elle pénalise les élèves quelque que soit leur appartenance sociale.
R. Goigoux (2000) Apprendre à lire à l'école : les limites d'une approche idéovisuelle. Psychologie française, N°45-3, 2000,233-243.

Exercice 2.

Il est admis que les conduites adaptatives des enfants lors de leur première entrée à la maternelle sont influencées par des facteurs différentiels tels que l'âge, le sexe, la durée de la journée scolaire. On étudiera ici le rôle des expériences antérieures de séparation et celui des habitudes de vie en collectivité à partir des observations comportementales de trois groupes d'enfants suivis durant leur première année de maternelle, dans la classe et dans la cour et différenciés selon leur mode de garde antérieur (mère, assistante maternelle, crèche). L'analyse statistique des modalités comportementales envers les enseignants, les pairs, le matériel, fait apparaître une différence qualitative et surtout quantitative entre les trois groupes; elle est surtout marqué dans le champ des comportements sociaux des enfants avec leurs pairs, puis dans l'intérêt pour le matériel. Les effets différentiels des modes de garde sont particulièrement nets au début de l'année scolaire et dans la cour de récréation. Ils ont tendance à se maintenir au terme d'une année scolaire et ce sont les enfants qui n'ont connu que la garde au foyer qui ont le plus de difficultés.
C. Sitbon-Zwobada (1997) Influence des modes de garde préscolaire lors de l'entrée en maternelle. Revue de psychologie de l'éducation, 1, 57-80.

---

[2] L'approche idéo-visuelle, encore appelée méthode globale, consiste à aborder la lecture par la reconnaissance de la forme visuelle des mots.
[3] L'approche phonique correspond à ce que d'autres auteurs appellent la méthode analytique. L'apprentissage de la lecture y est abordé par l'identification des sons composant le mot.
[4] Dans ce contexte, didactique signifie « méthode d'enseignement ».
[5] Le code grapho-phonologique est la correspondance entre l'écriture et la prononciation des syllabes ou des mots.

Exercice 3

Le but de cette étude est d'explorer l'influence de l'affect sur la résolution de problème par analogie. Quatre affects spécifiques, positif, neutre triste et agressif, sont induits à partir d'extraits musicaux et d'items d'imagerie. Les sujets doivent résoudre soit des problèmes bien définis, soit des problèmes mal définis. Le raisonnement analogique implique d'utiliser l'information source avant l'induction de l'humeur afin de résoudre un problème cible. Un groupe Le contrôle ne reçoit pas d'information source. Une performance accrue est attendue en condition « humeur positive » pour le problème mal défini alors qu'elle devrait être altérée pour les problème bien définis.

P. Cabrol (1998) Induction d'humeur et résolution de problème par analogie. VIIe colloque de l'association pour la recherche cognitive, Arc'98, 11 et 12 Décembre 1998, Université de Paris 8, Saint-Denis.

RECODAGE DE VARIABLE.

RÉSUMÉ

Lorsque sous leur forme originale, les données ne sont pas utilisables, ou lorsque le nombre d'observations n'est pas au moins égal à 5 fois le nombre de modalités de la variable, il faut procéder à un recodage de la variable.
Le recodage est applicable à tous types d'échelles de mesure.
On peut recoder une variable par regroupement de modalités. Dans ce cas, le recodage n'est pas réversible.
Dans le cas d'une variable nominale, on regroupera les modalités par proximité de sens en les désignant par une notion plus générale.
Dans le cas d'une variable ordinale, on regroupera les modalités voisines en les désignant par une notion plus générale.
Dans le cas d'une variable d'intervalle ou numérique, on regroupe les modalités voisines en les remplaçant par la valeur centrale de la classe, c'est-à-dire la moyenne des valeurs appartenant à la classe.
On peut également recoder une variable par transformation. Cette transformation n'est rien d'autre que l'application d'une fonction mathématique aux données. Dans ce dernier cas, le recodage est réversible.

LES PIÈGES À ÉVITER.

Constituer des classes non exhaustives : Lorsqu'on recode les variables par regroupement, il est essentiel de veiller à ce que toutes les observations trouvent une place dans les classes.
Constituer des classes non exclusives : dans un recodage par regroupement de modalités, il faut que les classes soient exclusives pour qu'on puisse classer les observations sans ambiguïté.

S'ENTRAINER.

Exercice 4. Dans certaines interfaces de programmation dédiées à l'expérimentation, l'unité de temps utilisé par la machine est de 1/60e de seconde. Dans une expérience, les temps de réponses observés vont de 21 à 55 soixantièmes de seconde. Recodez la variable pour exprimer ces temps en millisecondes.

Exercice 5. Sachant que pour l'expérience évoquée à l'exercice 4, nous avons 40 observations, réalisez un regroupement en classes en conservant un maximum de classes sans étendre l'intervalle de variation.

RESUMER UN PROTOCOLE: LA DISTRIBUTION

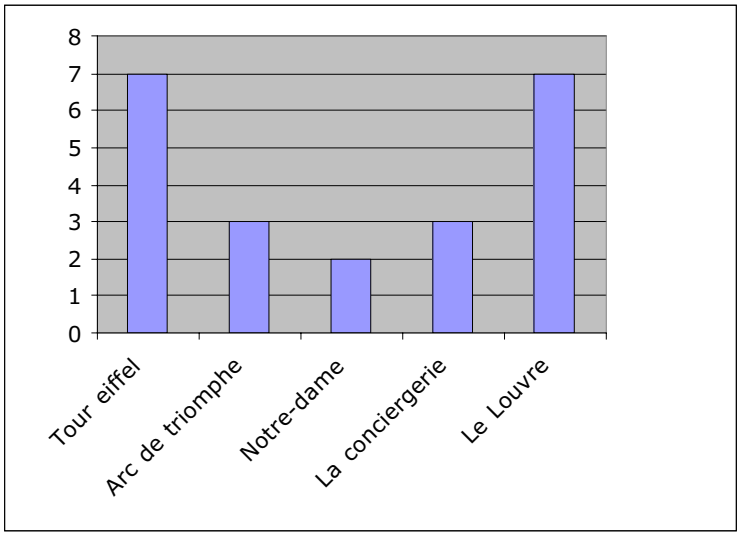
RÉSUMÉ

Une première façon de résumer un protocole est de faire une distribution, c'est-à-dire le dénombrement des observations pour chacune des modalités.
Le nombre d'observation pour chaque modalité est appelé effectif. Il est noté n(u). Le nombre total d'observation est noté n.
En divisant l'effectif par le nombre total d'observation on obtient la fréquence. La fréquence d'une modalité est notée f(u). $f(u)=n(u)/n$
En multipliant la fréquence par 100, on obtient un pourcentage.
Pour les protocoles qui sont recodés en classes, on peut calculer pour chaque classe sa densité d'effectifs qui est égale à l'effectif divisé par le nombre de modalités regroupées dans la classe.
On peut également résumer un protocole par une distribution cumulée à gauche (ordre croissant) ou à droite (ordre décroissant).
On peut également représenter une distribution, cumulée ou non par un graphique (diagramme en bâtonnets, en secteur ou histogramme).

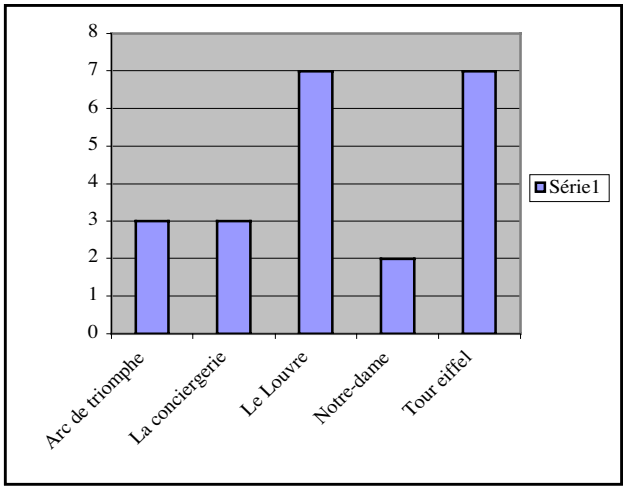
LES PIÈGES À ÉVITER.

Oublier les modalités pour lesquelles il n'y a aucune observation. C'est une erreur fréquente, surtout si on n'a pas pris la précaution de lister l'ensemble des modalités.

Commenter, dans un graphique, la forme d'une distribution sur une variable nominale. On ne peut rien dire de la forme d'une telle distribution. En effet, il n'y a pas de contrainte sur l'ordre des modalités avec ces variables. En réarrangeant les modalités, on peut obtenir une distribution qui a une autre allure. Imaginons par exemple qu'on demande à des touristes de passages dans la capitale quel est le monument qu'ils ont préféré. La distribution (imaginaire) des réponses est la suivante. Nous la représentons sous forme graphique.



Sur un tel graphique, on est tenté de voir une distribution en cloche inversée, en apparence symétrique. Mais comme il n'y a pas de raison pour ranger les modalités de cette façon, on peut tout aussi bien les classer par ordre alphabétique. Le graphique de distribution devient alors le suivant :

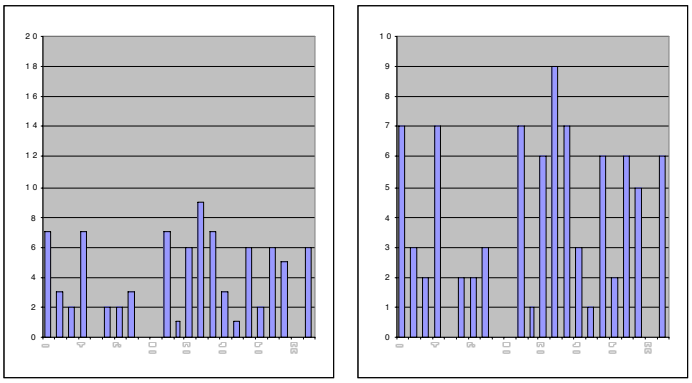


Confondre une distribution et une distribution cumulée. Cela arrive lorsqu'il est demandé de travailler sur les deux types de distribution simultanément, et notamment sous forme graphique. S'ensuivent alors des commentaires du genre "c'est la dernière modalité qui constitue le mode" où "les effectifs augmentent en même temps que les modalités. Dans une distribution cumulée à gauche, l'effectif cumulé de la dernière modalité est toujours le plus important et les effectifs cumulés à gauche croissent toujours dans le même sens que les modalités de la variable. Pour éviter ce genre de piège, il faut jeter un petit coup d'œil sur l'échelle des effectifs, si l'effectif maximum correspond au nombre de sujets, c'est que vous avez devant vous le graphique de la distribution cumulée.

Orienter l'axe de la variable quand celle-ci est nominale ou que la liste des modalités est exhaustive. Orienter un axe, c'est le terminer par une petite flèche. Celle-ci signale que les modalités sont ordonnées (ce qui n'est pas le cas des variables nominales, elle n'a donc

dans ce cas pas de sens) et que la liste des modalités portées sur le graphique n'est pas exhaustive (ce qui est souvent le cas des variables numériques pour lesquelles il peut exister d'autres modalités observables au-delà des modalités observées). La plupart des logiciels de type "tableur" ignorent purement et simplement l'orientation de l'axe. Cette erreur est donc surtout faite lorsqu'on doit faire le graphique à la main (durant les examens par exemple) alors avant de mettre la petite flèche qui semble tant améliorer l'esthétique du graphique, il faut s'interroger sur son sens.

Choisir une échelle des effectifs inappropriée. L'échelle de mesure doit être en rapport avec l'ordre de grandeur des effectifs. En choisissant une échelle de mesure trop petite, on a tendance à écraser le graphique et à minimiser des différences qui pourtant existent. À l'inverse, une échelle de mesure trop grande peut faire croire à des différences qui en réalité sont minimes. À titre d'exemple, Comparez ces deux graphiques réalisés sur une même distribution.



Le premier donne l'impression de différences d'effectifs peu importantes. Si on fait la même chose, mais en prenant une échelle plus petite pour les effectifs, on a au contraire l'impression de différences plus importantes.

S'ENTRAINER.

Exercice 6.

À titre d'exemple, voici un protocole correspondant à la passation, par un ensemble de 113 sujets d'un test (Faverge, 1966). Les individus statistiques sont les sujets. La variable correspond au nombre de réponses correctes sur un ensemble de 50 items. Nous avons une seule variable, l'échelle de mesure est une échelle d'intervalle. Nous avons donc ici un protocole univarié non structuré. Réalisez la distribution de cette variable.

I	Note	I	Note	I	Note	I	Note	I	Note	I	Note	I	Note	I	Note	I	Note
S1	43	S14	40	S27	38	S40	33	S53	32	S66	37	S79	44	S92	42	S105	27
S2	31	S15	48	S28	33	S41	33	S54	41	S67	39	S80	19	S93	44	S106	33
S3	38	S16	37	S29	44	S42	35	S55	31	S68	36	S81	34	S94	20	S107	25
S4	40	S17	40	S30	44	S43	45	S56	36	S69	10	S82	39	S95	44	S108	18
S5	40	S18	32	S31	49	S44	43	S57	29	S70	26	S83	24	S96	21	S109	44
S6	41	S19	13	S32	24	S45	9	S58	22	S71	23	S84	34	S97	26	S110	37
S7	28	S20	24	S33	18	S46	36	S59	36	S72	27	S85	34	S98	32	S111	48
S8	41	S21	28	S34	49	S47	30	S60	37	S73	26	S86	21	S99	36	S112	29
S9	43	S22	33	S35	41	S48	12	S61	35	S74	20	S87	38	S100	44	S113	35
S10	41	S23	18	S36	25	S49	25	S62	30	S75	24	S88	19	S101	37		
S11	41	S24	36	S37	27	S50	16	S63	25	S76	37	S89	42	S102	8		
S12	41	S25	47	S38	43	S51	36	S64	20	S77	45	S90	46	S103	29		
S13	35	S26	29	S39	11	S52	25	S65	39	S78	43	S91	50	S104	42		

Exercice 7.

Représenter graphiquement la distribution du protocole de l'exercice précédent.

Exercice 8.

Faire la distribution des notes au test en 7 classes d'intervalles égaux et en s'arrangeant pour que la valeur extrême de la dernière classe soit égale à la modalité observée la plus élevée.

Exercice 9.

Même exercice que le 8, mais avec 9 classes.

RESUMER UNE DISTRIBUTION : LES PRINCIPAUX INDICES STATISTIQUES.

RÉSUMÉ.

Pour résumer une distribution, on calcule des indices de position ou de tendance centrale et des indices de dispersion. Le choix des indices dépend de ce qu'on souhaite résumer dans la distribution et de l'échelle de mesure de la variable (se reporter au tableau suivant).
Les questions qu'on peut se poser sont :

- Sur quelle(s) modalité(s) se concentre(nt) les observations (concentration) ?
- Comment les observations se répartissent-elles dans la distribution (répartition) ?
- Quel est le centre de gravité de la distribution et la variation moyenne autour de ce centre (centre et variation) ?

		Echelles de mesure		
Questions	**Indices**	**nominale**	**Ordinale**	**Numérique**
Concentration	de position	mode	mode	mode
		Mode secondaire	Mode secondaire	Mode secondaire
Répartition	de tendance centrale		Médiane	Médiane
	de dispersion		Quartile 1 et 3	Quartile 1 et 3
Centre et variation	de tendance centrale			Moyenne
	de dispersion			Ecart-type

Le mode est la modalité la plus fréquente.
Le mode secondaire est la modalité la plus fréquente après le mode.

La médiane (ou quartile 2) est la modalité dont l'effectif cumulé correspond à n/2.
Le quartile 1 est la modalité dont l'effectif cumulé à gauche correspond à n/4.
Le quartile 3 est la modalité dont l'effectif cumulé à gauche correspond à n*3/4.

La moyenne est la somme des observations divisée par le nombre d'observations. On la note m. Elle représente le centre de gravité de la distribution : $m=\Sigma x/n$
La variance est la moyenne des carrés des écarts à la moyenne. On la note s^2 :

$$s^2 = \frac{\sum (x_i - m)^2}{n} = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n}$$

L'écart-type est la racine carrée de la variance. On le note s. Il représente la variation moyenne des observations autour de la moyenne du protocole. On l'utilise comme indice de dispersion pour résumer un protocole.

La variance corrigée est la somme des carrés des écarts à la moyenne pondérée par n-1. On la note s^2_{corr} = var_{corr} :

$$s^2_{corr} = \frac{\sum (x_i - m)^2}{n-1} = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} = \frac{s^2 * n}{n-1}$$

L'écart-type corrigé est la racine carrée de la variance corrigée. Il représente d'un point de vue descriptif une mesure de la variabilité intersujet, et d'un point de vue inférentiel, une estimation de l'écart-type de la population parente. On l'utilise lorsqu 'on veut essayer de généraliser les observations faites sur l'échantillon à l'ensemble de la population dont il est issu.

LES PIÈGES À ÉVITER.

Déclarer un effectif cumulé au lieu d'une modalité comme médiane ou quartile. C'est une erreur assez fréquente qui consiste à annoncer que la médiane est égale à n/2 au lieu d'annoncer la modalité correspondante. Pour l'éviter, il faut garder à l'esprit que médiane et quartiles sont des coupures de l'échelle de mesure (et non de la distribution cumulée). Ce qu'on désire savoir avec ces indices c'est sur quelles modalités se répartissent les observations et non la valeur de n/2.

Confondre la somme des observations avec la somme des modalités. Cette erreur très fréquente résulte d'une confusion entre le tableau de protocole et le tableau de distribution. Il est donc tout à fait essentiel de bien les différencier. Le moyen le plus efficace pour faire la différence est de vous demander à quoi correspondent les lignes du tableau. Si les lignes correspondent aux individus statistiques alors vous travaillez sur le tableau du protocole. Si les lignes correspondent aux modalités de la variable alors vous travaillez sur le tableau de distribution. Pour vous y retrouver, il est impératif d'avoir au préalable identifié vos unités statistiques et vos variables.

Confondre le nombre d'observations et le nombre de modalités. Cette erreur tout aussi fréquente que la précédente relève des mêmes difficultés et sera évitée de la même façon.

Confondre le carré de la somme des observations et la somme des carrés des observations. Cette erreur est liée entre autres à une difficulté à intégrer les écritures formelles et à un non-respect de l'ordre de priorité des opérations. Le carré de la somme des observations $(\Sigma x)^2$ suppose qu'on fasse d'abord la somme des observations (souvenez-vous qu'il faut d'abord traiter l'intérieur des parenthèses). La somme des carrés des observations Σx^2 requiert de faire d'abord l'élévation au carré, puis la somme (l'ordre de priorité des opérations n'a pas changé depuis que vous avez quitté le collège, on commence d'abord par

les puissances, on fait ensuite les multiplications ou les divisions et on termine par les sommes et les différences).

Annoncer une variance ou un écart-type négatif. La variance est une moyenne de carrés. Un carré étant forcément positif (la multiplication de deux nombres relatifs de même signe donne un nombre positif), leur moyenne ne saurait être négative. Elle représente une distance et une distance négative n'a pas de sens.

Annoncer une moyenne qui sort de l'intervalle de variation de la variable. Cette erreur, pour le moins étonnante, consiste a annoncer une moyenne qui est située au delà de la plus petite ou de la plus grande des observations. La moyenne est le centre de la distribution, elle est donc forcément située entre les deux extrêmes observés.

Annoncer un écart-type plus grand que la moyenne. Cette erreur résulte d'une mauvaise représentation de ce que sont un écart-type et une moyenne. L'écart-type représente la distance moyenne qui sépare les observations de la moyenne du protocole, la moyenne étant le centre de la distribution de ce protocole. La plus grande distance qu'on puisse trouver entre la moyenne et les deux extrémités de la distribution, c'est lorsque la moyenne est au milieu de l'intervalle de variation. L'écart moyen entre les observations et cette moyenne est alors au maximum de la moitié de l'intervalle de variation. L'écart-type est donc forcément inférieur ou égal à la moyenne (sauf si votre échelle de mesure comprend des nombres négatifs). Le cas d'égalité entre la moyenne et l'écart-type est un cas extrême et rarissime. Il suppose que la moyenne soit située à mi-chemin entre la plus grande et la plus petite valeur observée et que toutes les valeurs observées soient situées uniquement sur les extrémités de la distribution. Dans ce cas, on peut douter du caractère numérique de la variable (quel est le sens de la notion d'intervalle si seulement deux modalités sont observées ?) et donc de la pertinence du calcul de la moyenne et de l'écart-type.

Confondre la moyenne du protocole avec la moyenne scolaire. Cette erreur consiste à prendre comme moyenne 10 au lieu de la moyenne du protocole, erreur surtout observée lorsque la variable va de 0 à 20. Le langage courant, sous l'effet de notre passage à tous dans le milieu scolaire, désigne sous le nom de moyenne la note 10. C'est effectivement une moyenne, au sens arithmétique du terme. C'est la moyenne des modalités de la variable (pour une note sur 20). Il ne faut, bien entendu, pas la confondre avec la moyenne des observations dont nous nous occupons en statistiques.

S'ENTRAINER.

Exercice 10.
Situez le mode de la distribution en classes sur la note au test (exercice du paragraphe précédent).

Exercice 11.
Situez la médiane et les quartiles de la distribution après regroupement en 9 classes des notes au test (exercice 9).

Exercice 12.
Calculer la moyenne et l'écart-type de la distribution des notes au test (exercice du paragraphe précédent).

SITUER UN INDIVIDU DANS UNE DISTRIBUTION

RÉSUMÉ.

Pour situer un sujet dans une distribution, il faut des repères. Ces repères sont :
- Des coupures spécifiques (déciles ou centiles) résumant la distribution.
- La moyenne et l'écart-type (note z ou note réduite).
On peut aussi combiner ces deux repères (normalisation).

Décilage et centilage.
L'échelle de mesure doit être ordinale ou numérique.
9 déciles partagent la distribution en 10 interdéciles comprenant 10% des observations.
99 centiles partagent la distribution en 100 intercentiles comprenant 1 % des observations.
La procédure de décilage ou de centilage est similaire à la procédure permettant la détermination des quartiles.

Note z ou note réduite.
La note z est égale à l'écart à la moyenne pondérée par l'écart-type. $z = (x_i-m) / s$
Elle revient à exprimer l'écart à la moyenne en nombre d'écart-type. La moyenne des notes z est égale à 0. L'écart-type des notes z est égal à 1.

Normalisation.
Une distribution normale est une distribution en forme de cloche centrée sur le mode, la médiane et la moyenne. Dans une distribution normale, connaissant la moyenne et l'écart-type, on peut déterminer la fréquence des observations pour chacune des modalités.
La normalisation est un recodage par regroupement de modalités destiné à modifier la forme de la distribution pour la centrer sur 0 (moyenne des notes z) et en faire une distribution normale. L'échelle de mesure doit être numérique. La procédure de normalisation est la suivante :

- Détermination des limites de classes en note z (4 / (le nombre de classe-1)
- Lecture des fréquences cumulées à gauche p(z<u) dans la table du z
- Calcul des effectifs cumulés : p(z<u)*n
- Détermination des coupures et des effectifs de chaque classe (procédure similaire au décilage).

LES PIÈGES À ÉVITER.

Chercher dix coupures dans le décilage. Cette erreur provient de la confusion entre les coupures (les déciles) et les intervalles délimités par ces coupures (les interdéciles). Pour éviter cette erreur, on peut faire l'analogie avec le partage d'un gâteau. Pour faire dix parts (les classes) , il faut donner 9 coups de couteau (les coupures).

S'ENTRAINER.

Exercice 13.
Effectuez un décilage sur la distribution des notes au test avant regroupement (paragraphes précédents). Situez par rapport à ce décilage un sujet qui aurait obtenu 25 au test.

Exercice 14.
Situez le même sujet ayant obtenu 25 au test à l'aide de la transformation en notes z.

Exercice 15.
Comment se situe ce même sujet ayant obtenu 25 dans la distribution normalisée des notes?

COMPARAISON A UNE DISTRIBUTION DE REFERENCE.

RÉSUMÉ.

La comparaison d'une distribution empirique à une distribution de référence répond à deux types d'objectifs :

Les données se répartissent-elles de façon aléatoire sur les différentes modalités ?
Dans ce cas, on présuppose que la distribution est théoriquement uniforme. On compare alors les données à une distribution plate en calculant la statistique X^2 (Khi-deux) donnée par la formule suivante :

$\chi^2 = \sum \frac{(obs - théo)^2}{théo}$ où *obs* est l'effectif observé et *théo*, l'effectif théorique. *Théo* est obtenu en divisant n par le nombre de modalités.

Vérifier les conditions d'application d'une procédure statistique,
Cela concerne principalement la comparaison d'une distribution empirique à une distribution normale (courbe de Gauss). Pour cela on calcule ce que devrait être la distribution si elle était normale (distribution théorique) puis on compare la distribution observée à la distribution théorique. Ce calcul s'effectue en quatre étapes :

A-. Calcul des notes z correspondant aux limites de chacune des classes.
B-. Recherche dans la table des proportions correspondant à notes z notée p(u<z).
C-. Calcul des effectifs cumulés théoriques : p(u<z)*n et des effectifs non-cumulés théoriques.
D-. Comparaison graphique de la distribution observée et de la distribution théorique.

LES PIEGES À EVITER.

Confondre les effectifs (ou fréquences) cumulés et les effectifs (ou fréquences) non cumulés. Dans la comparaison à une distribution normale, on doit à plusieurs reprises passer de l'un à l'autre, d'où le risque de les confondre. Pour éviter cela, on regarde si les effectifs augmentent dans le même sens que les modalités. Il est, en effet, très rare que les effectifs non cumulés soient croissants en même temps que la variable (ce qui est toujours le cas des effectifs cumulés).

Prendre la valeur centrale de la classe au lieu de la limite de classe pour le calcul des notes z. C'est une erreur relativement fréquente. En effet, pour calculer la moyenne, on prend la valeur centrale de la classe puisqu'on cherche la tendance centrale de la distribution. En revanche, pour le calcul des notes z, il faut prendre la limite de la classe, puisque la table nous donne des fréquences cumulées, c'est-à-dire la proportion d'observation inférieure ou supérieure (selon le sens du cumul) à une valeur donnée de z.

S'ENTRAINER.

Exercice 16.

Dans une expérience sur le raisonnement, on demande à un ensemble de 40 sujets de choisir parmi différentes conclusions de l'argument suivant celle qui convient :

Si j'étais riche alors je m'achèterais une nouvelle voiture.
Je me suis acheté une nouvelle voiture.
Donc…
A-. Je suis riche.
B-. Je ne suis pas riche.
C-. *On ne peut pas savoir.*

La distribution des sujets sur ces trois réponses possibles est donnée dans le tableau suivant :

Conclusions	*Effectifs.*
Je suis riche.	25
Je ne suis pas riche.	20
On ne peut pas savoir.	15
Total	60

Peut-on dire que les sujets répondent au hasard ?

Exercice 17.

Dans cet exercice, vous reprendrez la distribution des notes au test en 9 classes utilisée dans les paragraphes précédents. Peut-on dire que cette distribution est très différente d'une distribution normale ? Pour plus de commodités, nous vous rappelons cette distribution ci-dessous.

Intervalle	*Valeur centrale*	*Limites*	*Effectifs*
6-10	8	5,5	3
11-15	13	10,5	3
16-20	18	15,5	9
21-25	23	20,5	13
26-30	28	25,5	14
31-35	33	30,5	17
36-40	38	35,5	23
41-45	43	40,5	24
46-50	48	45,5	7
		Total	113

COMPARAISON DE GROUPES D'OBSERVATIONS.

RÉSUMÉ.

Pour effectuer une comparaison de groupes d'observations, il faut déterminer les groupes qui vont être comparés et la base de cette comparaison, c'est-à-dire sur quoi vont être comparés les groupes. Il faut au minimum un protocole structuré univarié structuré, c'est-à-dire disposer d'au moins un facteur (variable indépendante) et d'une variable observée (variable dépendante).

- ✓ La détermination des groupes d'observations dépend de la structure du protocole (emboîtement ou croisement) c'est-à-dire de la relation entre le facteur « sujet » et la variable indépendante.

- ✓ Le choix de la base de comparaison dépend de l'échelle de mesure (numérique, ordinale ou nominale) de la variable dépendante. Très généralement cette base de comparaison est un résumé de la distribution (indice de position ou de tendance centrale) combiné ou non avec un indice de dispersion.

Les bases de comparaison sont généralement les moyennes, les médianes et les fréquences. Cette liste n'est pas exhaustive et d'autres indices peuvent être utilisés en fonction de la nature des données. Le tableau ci-dessous indique pour quelle échelle de mesure, elles sont utilisables. La procédure de comparaison consiste à calculer ces indices pour chacun des groupes d'observations et ensuite à les comparer.

Echelle de mesure	Moyenne	Médiane	Fréquence
Numérique	Oui	Oui	Oui
Ordinale		Oui	Oui
Nominale			Oui

LES PIEGES À EVITER.

Calculer les fréquences sur le total général. Le principe de la comparaison de groupes est de réitérer l'analyse descriptive pour chacune des modalités du facteur. Le calcul des fréquences sur le total général ne permet pas de comparer les groupes. Nous verrons dans le chapitre suivant que cela correspond plutôt à une étude de la liaison entre variables.

Calculer les fréquences sur les totaux correspondant aux modalités de la variable dépendante. Dans ce cas, les groupes d'observations sont les modalités de la variable dépendante, et ce sont elles que vous comparez, ce qui n'est, bien entendu, pas équivalent à la comparaison des groupes sur les modalités du facteur. Cette erreur résulte du fait que les données sont généralement résumées sous la forme d'un tableau à double entrée croisant les modalités du facteur et de la variable observée. Il faut alors considérer le tableau soit en ligne soit en colonne selon le sens dans lequel il est disposé.

Repérer les quartiles en prenant, pour n, le nombre total d'observations et non le total des observations de chaque groupe. Cette erreur est de même nature que celle qui consiste à calculer les fréquences sur le total général.

S'ENTRAINER.

Exercice 18.
Lors d'une enquête sur les conditions de travail dans un centre hospitalier spécialisé, on a posé au personnel la question suivante : « Pensez-vous que l'emploi soit menacé dans le milieu hospitalier ? ». Deux catégories de personnel ont été interrogées, ceux qui travaillent à l'hôpital (Intra) et ceux travaillant sur dans des structure extra-hospitalière (extra). Voici la distribution des réponses. Réalisez une analyse descriptive de ces données pour comparer les deux groupes de personnel. Commentez.

	Intra	*Extra*	*Total*
Oui	150	29	179
Non	116	43	159
Pas de réponse	23	7	30
Total	289	79	368

Exercice 19.

Dans une enquête sur la perception des causes d'accidents de la route, on a posé la question suivante : Pensez-vous que les défaillances mécaniques sont une source d'accidents ?

Les sujets avaient à répondre sur une échelle allant de 0 à 10 dans laquelle la note 0 correspond à « jamais », la note 5 à « parfois » et la note 10 à « toujours ». Sur les 160 sujets interrogés, voici la distribution des réponses pour les hommes et les femmes. Réalisez une analyse descriptive de ces données pour comparer les réponses à cette question en fonction du sexe. Commentez.

Jugement	*Notes*	*Femmes*	*Hommes*
Jamais	0	1	2
	1	3	2
	2	8	4
	3	10	4
	4	9	8
Parfois	5	12	12
	6	12	13
	7	8	11
	8	9	10
	9	7	9
Toujours	10	1	5
		80	80

Exercice 20.

Lors d'une enquête, on relève**,** sur les registres d'état civil, l'âge de l'époux et l'âge de l'épouse au moment du mariage pour 39 couples. On se demande si les hommes et les femmes se marient en général au même âge. Les données sont présentes ci-dessous sous la forme de couples de chiffres dont le premier est l'âge de l'époux et le second l'âge de l'épouse :

(20;20), (25;24), (25;22), (31;29), (18;20), (25;24), (32;31), (26;25), (22;21), (24;24), (24;22), (25;26), (27;22), (20;18), (18;19), (23;21), (23;23), (34;35), (45;43), (23;22), (22;21), (20;22), (21;22), (34;32), (21;20), (32;30), (45;44), (33;32), (31;30), (33;31), (36;35), (28;27), (27;26), (29;28), (21;21), (22;20), (24;23), (31;29), (33;30),

Peut-on dire que les hommes se marient plus tard que les femmes ? Réalisez une analyse descriptive des données pour répondre à cette question. Commentez.

ÉTUDE DE LA LIAISON ENTRE DEUX VARIABLES

RÉSUMÉ.

Pour étudier la liaison entre deux variables, il faut un protocole bivarié, structuré ou non, c'est-à-dire disposer de deux variables observées. La question posée n'est pas alors de savoir si les distributions des groupes d'observations sont similaires ou non pour chacune des modalités d'un facteur, mais si la connaissance de la modalité observée sur une des variables pour un individu statistique permet ou non de dire quelque chose de la modalité qu'on devrait observer sur l'autre variable. On cherche alors à prédire ce qu'on devrait observer sur une variable à partir de ce qu'on sait sur l'autre variable. La procédure d'analyse à employer dépendant de l'échelle de mesure des variables.

- Dans le cas des variables nominales, on analysera la distribution des effectifs conjoints en calculant le carré moyen de contingence Φ^2 (lire Phi-deux). Cet indice varie entre 0 et 1 et représente la proportion d'observations observées qui s'écartent de la distribution théorique qu'on obtiendrait si la liaison entre les variables était nulle. 0 où une valeur proche de 0 signifie que les variables sont peu ou pas liées. 1 ou une valeur proche de 1 signifie que les variables sont fortement liées. On est alors proche du cas de concordance, c'est-à-dire qu'à chaque modalité de première variable correspond une seule modalité de la seconde variable. Ce dernier cas autorise donc avec une marge de sécurité satisfaisante de prédire la valeur que prendrait une variable à partir de l'autre.

$$\phi^2 = \sum \frac{(n_{jk} - n\copyright_{jk})^2}{(n\copyright jk)^2} * \frac{n\copyright jk}{n}$$

- Dans le cas des variables numériques, on analyse la relation entre les variables en calculant le r de Bravais-Pearson. Cet indice varie entre –1 et 1. La valeur absolue de cet indice indique la part de la variance expliquée par une fonction linéaire : égal ou proche de 0, la liaison est nulle ; égal ou proche de 1, il existe une relation de proportionnalité entre les variables. Autrement dit on peut prédire une variable à partir d'une autre à l'aide d'une équation de type y=ax+b, où y est la variable à prédire et x la variable connue. Graphiquement, les données s'alignent selon une droite. Le signe de cet indice indique le sens de la liaison : liaison inverse pour les valeurs négatives, variation dans le même sens pour les valeurs positives.

$$r = \frac{\sum (x_i - m_x)(y_i - m_y)}{\sqrt{\sum (x_i - m_x)^2 \sum (y_i - m_y)^2}}$$

- Dans le cas des variables ordinales ou numériques lorsque le r de Bravais-Pearson n'est pas une bonne mesure de la corrélation**,** c'est-à-dire lorsque graphiquement la relation entre les variables est clairement non-linéaire, on analyse la covariation des variables en calculant le coefficient de corrélation par rang de Spearman ρ (lire r de Spearman). Cet indice varie entre –1 et 1. Comme le r de Bravais-Pearson dont il dérive, le r de Spearman indique la part de la variance expliquée par une fonction linéaire. Le signe de cet indice indique le sens de la liaison. Il s'interprète donc de la même façon.

$$\rho = 1 - \frac{6\Sigma d^2}{n^3 - n}$$

LES PIEGES À EVITER.

Calculer les fréquences sur l'effectif total en ligne ou en colonne. Attention, contrairement à la comparaison de groupes d'observations, l'évaluation de la relation entre deux variables nominales nécessite un calcul des fréquences sur le total des observations et non sur les totaux marginaux.

S'ENTRAINER.

Exercice 21.

Pour cet exercice, nous reprendrons les données de l'enquête sur la représentation de la psychologie (voir les annexes de votre cours). Nous nous poserons la question de la relation entre la profession et la réponse à la question I : « Pourquoi, à votre avis les gens vont-ils voir les psychologues ? ». Nous donnons ci-dessous la distribution des effectifs conjoints (encore appelés effectifs observés).

Effectifs conjoints	*Cadre sup.*	*Cadre moy.*	*Ouvriers*	*Prof. Lib.*	*Total*
Parce qu'ils se sentent dans un état anormal.	1	7	3	6	17
Par besoin d'aide, de conseil.	10	6	2	6	24
Par besoin de se connaître.	6	3	8	3	20
Pour des problèmes d'orientation.	2	1	3	5	11
Autres réponses ou absence de réponses.	1	3	4	0	8
Total	20	20	20	20	80

Exercice 22.

Pour cet exercice, on reprendra les données de l'enquête sur les registres d'état civil (exercice 3 du chapitre précédent). La question qu'on se pose maintenant est : « existe-il une relation entre l'âge de l'époux et l'âge de l'épouse au moment du mariage ? ».

LA FOIRE AUX QUESTIONS.

Voici, en vrac, quelques-unes des questions les plus fréquemment posées. Si vous ne trouvez pas de réponse à votre question n'hésitez pas à contacter les enseignants lors de leur permanence.

Comment arrive-t-on à déterminer le nombre de classes dans un recodage de variable ?
Il n'y a pas de règle absolue pour déterminer le nombre de classes à prendre, cela dépend du nombre de modalités de la variable dans le protocole de base et du niveau de finesse dans la mesure dont le chercheur a besoin pour répondre aux questions qu'il se pose. Le nombre de classes dans un recodage est un compromis entre un nombre de modalités suffisamment importantes pour que la distribution soit informative et un nombre de modalités qui soit suffisamment faible pour que le protocole soit manipulable. Il faut garder à l'esprit que si le nombre de modalités est trop important, la distribution ne présentera pas d'intérêt. Par ailleurs chaque fois que je fais des regroupements de modalités, je perds de l'information. En exagérant, on pourrait même avoir des protocoles pour lesquels on n'aurait qu'une observation par modalité, ce qui est presque le cas dans l'exemple du " test des labyrinthes " puisque nous avons 74 observations pour 71 modalités observées.

À partir de quels éléments définit-on l'intervalle des classes ?
L'intervalle des classes est le nombre de modalités contenues dans la classe. Il doit être le même pour toutes les classes si on veut garder la structure d'intervalle de la variable dans le protocole de base. Cet intervalle est calculé en faisant le rapport entre le nombre de modalités observées pour la variable dans le protocole de base et le nombre de classes retenu. Dans l'exemple du cours, les valeurs extrêmes observées sont 26 et 97 secondes. Ce qui nous donne une étendue de: 97-26= 71 modalités observées. On désire ne garder qu'une dizaine de classes. L'intervalle est donc de :71/10=7,1. Ce qui nous donne un intervalle de 7. Si on ne prenait que 10 classes une des modalités observées ne serait pas prises en compte (7*10=70). Nous ne respecterions pas la contrainte d'exhaustivité des classes. Cela nous conduit à retenir 11 classes.

Quelle calculette conseillez-vous pour les examens de statistiques ?
Comme vous avez pu le lire dans le règlement d'examen, les calculatrices programmables sont interdites, même si l'étudiant déclare ne pas savoir se servir de la programmation. Pour vos études et l'examen, n'importe quelle calculette non programmable peut faire l'affaire, mais le plus judicieux est de s'assurer qu'elle intègre quelques fonctions statistiques comme la somme des carrés, la somme des observations, voir l'écart-type.

Est-ce que les erreurs de calculs comptent beaucoup dans les examens de statistiques ?
Bien sûr ce que nous attendons des étudiants c'est qu'il intègrent la démarche intellectuelle de la recherche. Les erreurs de calculs ont donc moins d'importance que les erreurs conceptuelles. Cependant, la bonne compréhension de la démarche d'analyse des données suppose l'acquisition d'un certain nombre de stratégie de contrôle sur ce qu'on fait (comprendre par exemple que la moyenne ne saurait sortir de l'intervalle de variation de la variable ou encore qu'une somme des carrés est forcément positive). Par ailleurs une erreur de calcul peut, et c'est souvent le cas des erreurs importantes, entraîner une erreur dans la conclusion.

Dans une distribution normale, comme celle de la variable z, la moyenne et l'écart-type sont-ils toujours égal à 0 et 1 ?
Dans une distribution normale, la moyenne et l'écart-type ne sont pas toujours égaux à 0 et 1. En revanche, pour la distribution de la variable z qui est un cas particulier de distribution normale, la moyenne est toujours égale à 0 et l'écart-type est toujours égal à 1.

Peut-on toujours calculer la distribution z?
On peut, par transformation de variable, ramener n'importe quelle distribution normale à une distribution z, si on connaît sa moyenne et son écart-type. Il faut donc au minimum que la variable soit numérique.

Pour calculer la densité, doit-on d'abord faire des regroupements de modalités ?
Non, lorsque la classe ne contient qu'une seule modalité, son étendue est de 1, et la densité est égale à la fréquence.

Page 591, pour le trapèze, n'est-ce pas la base, et non la hauteur qui a la valeur 0,50 ?
Par définition, un trapèze est un quadrilatère possédant deux cotés parallèles et sa hauteur est la distance entre ces deux droites parallèles. 0,5 est donc bien la hauteur du trapèze.

Qu'est-ce que p(z<u)?
p est une proportion qu'on interprète parfois en termes de probabilités p(z<u) est la proportion d'observations pour lesquelles z<u. C'est la fréquence cumulée à gauche des notes z< u. En termes de probabilités, c'est la probabilité que z<u. On appelle u le z théorique, c'est-à-dire celui de la table.

Par quel calcul, en utilisant quels chiffres, arrive-t-on aux résultats 0.16 et 0.31 à la page 592?

Les valeurs utilisées p 592 ne sont pas calculées, mais lues dans la table de la page 594. ces valeurs ont été arrondies. Pour z=-1, on lit dans la table 0,159 soit environ 0,16. Pour z=-0,5, on lit dans la table 0,309 soit environ 0,31

Pourquoi calcule-t-on les taux de liaison à partir des effectifs et non des fréquences (on gagnerait le calcul des effectifs théoriques...) ?

Dans le calcul du X^2, on ne peut faire l'économie du calcul des effectifs théoriques puisque par définition X^2 est la somme des carrés des différences entre les effectifs observés et les effectifs théoriques pondérés par les effectifs théoriques. On peut les calculer à partir des effectifs marginaux en appliquant une simple règle de trois (ligne* colonne/ effectif total), ce qui revient indirectement à calculer les fréquences.

A quoi servent les indices de corrélation ?

Les indicateurs de corrélation comme le r de Bravais-Pearson ou le carré moyen de contingence servent à évaluer le lien entre deux variables.

Comment calcule-t-on les limites inférieures et supérieures dans un regroupement par classes ?

Les limites de classes sont calculées à partir des valeurs centrales des classes. Pour la limite inférieure, on additionne la valeur centrale de la classe et la valeur centrale de la classe précédente. On divise ensuite par 2. Pour la limite supérieure, on additionne la valeur centrale de la classe et la valeur centrale de la classe suivante. On divise ensuite par 2. Notez au passage que la limite inférieure d'une classe est la limite supérieure de la classe précédente. De même que la limite supérieure d'une classe est la limite inférieure de la classe suivante.

Quelle est la différence entre une variable ordinale et une variable numérique ?

Une variable ordinale est caractérisée par un ordre entre les modalités, mais la notion d'intervalle entre les modalités n'a pas de sens. On ne peut pas calculer de moyenne sur de telles données. Par exemple le niveau scolaire (primaire, secondaire, supérieur). Une variable numérique est caractérisée par un ordre et un intervalle. Cela a du sens de calculer une moyenne sur ce type de variable. C'est par exemple la taille ou le poids d'un individu.

A l'examen, est-il toujours nécessaire de détailler tous les calculs?

Il n'y a, bien entendu, pas d'obligation de donner tout le détail des calculs. Cependant, en cas d'erreur de calcul, il ne m'est pas possible de différencier les erreurs de conceptualisation et les simples erreurs de calculs. Par défaut, je considèrerai que ce sont des erreurs de conceptualisation (qui sont davantage pénalisées). Je vous recommande donc de donner dans votre copie suffisamment de détails pour me permettre de m'assurer que vous avez compris la démarche d'analyse.

Quand doit-on utiliser une normalisation, une comparaison à une distribution de référence, une comparaison à une distribution uniforme ou à une distribution normale ?
Dans la comparaison à la distribution de référence, vous avez deux possibilités, comparer la distribution empirique : à une distribution uniforme ou à une distribution normale. La comparaison à une distribution uniforme est utilisée lorsqu'on se demande s'il existe ou non des différences entre les effectifs des différentes modalités. Dans un questionnaire par exemple, on peut se demander si les sujets choisissent plus souvent ou non une des modalités de réponses. La comparaison à une distribution normale est utilisée lorsqu'on a besoin de décrire la distribution par les seuls paramètres que sont la moyenne et l'écart-type ou lorsqu'on veut pouvoir considérer la moyenne comme la mesure "vrai" et l'écart-type comme un estimateur des erreurs de mesure. On s'en sert également dans les inférences statistiques que vous verrez en deuxième année. La normalisation est une opération de transformation de variable destinée à ramener la distribution à une distribution aussi proche que possible d'une distribution normale.

Concernant le carré moyen de contingence, peut-on appliquer X^2 plutôt que Φ^2? Quelle est finalement la différence ?
Ces deux statistiques sont similaires, cependant, X^2 dépend de N, alors que ce n'est pas le cas de Φ^2, puisque $\Phi^2=X^2/N$. On préférera donc pour l'analyse descriptive Φ^2, tandis que X^2 sera réservé à l'analyse inférentielle que vous étudierez l'an prochain.

Comment calcule t-on l'intervalle de classe dans la normalisation ?
L'intervalle de classe est égal à l'étendue de la variable divisée par le nombre de coupures qu'on désire effectuer sur l'échelle de mesure. Comme avec les quartiles et les déciles, le nombre de coupures nécessaires est égal au nombre d'intervalles à obtenir moins un. L'étendue est de 4 (2 écart-types au-dessus et en dessous de la moyenne). Pour obtenir 11 classes, il nous faut 10 coupures. L'intervalle de classe est donc de 4/10=0,4. Pour obtenir 9 classes, il nous faut 8 coupures. L'intervalle de classe est donc de 4/8=0,5. Pour obtenir 7 classes, il nous faut 6 coupures. L'intervalle de classe est donc de 4/6=0,67 (valeur arrondie).

CORRIGÉS

Exercice 1.

Dans cette recherche, les individus statistiques sont les 76 enfants.

Le facteur étudié est la méthodologie didactique. C'est une variable nominale dichotomique (2 modalités). Les sujets ne voient qu'une des deux méthodes, on a donc une relation d'emboîtement.Les variables observées (variables dépendantes ou VD) sont :

- ✓ La performance initiale en fin de maternelle.
- ✓ Le score à l'évaluation nationale de CE2.
- ✓ La vitesse d'identification des mots.

Aucune précision n'est donnée sur la nature des mesures et les échelles de mesure retenues dans cette étude. Il s'agit d'un protocole multivarié structuré. L'objectif de cette recherche est la comparaison des deux méthodologies didactiques. Le tableau représentant le protocole serait alors le suivant :

Individus	*Méthodologie didactique*	*Performance initiale*	*Evaluation CE2*	*Vitesse d'identification des mots*
Sujet 1	Idéo-visuelle	VD 1	VD 2	VD 3
Sujet 2	Phonique	VD 1	VD 2	VD 3
Sujet 3 …	Etc.	VD 1	VD 2	VD 3

Exercice 2.

Dans cette expérience, les individus statistiques sont les enfants. Ils sont répartis dans trois groupes caractérisés par le mode de garde antérieure (on a donc une relation d'emboîtement entre ce facteur et le facteur sujet).

La variable observée est constituée par les modalités comportementales. Aucune précision n'est donnée sur l'échelle de mesure utilisée. Les modalités comportementales sont observées dans trois conditions (envers les enseignants, les pairs et le matériel). On a donc une relation de croisement. Le protocole est donc un protocole univarié structuré.

L'objectif de cette recherche est l'évaluation de la relation entre le mode de garde antérieure et les modalités comportementales à l'égard de l'environnement scolaire (enseignants, pairs et matériel). Le tableau représentant le protocole serait alors le suivant :

Individus	*Mode de garde*	*Comportement envers les enseignants*	*Comportement envers les autres enfants*	*Comportement envers le matériel*
Sujet 1	Mère.	VD	VD	VD
Sujet 2	Assistante maternelle.	VD	VD	VD
Sujet 3 …	Crèche…	VD	VD	VD

Exercice 3.

Dans cette expérience, les individus statistiques sont les sujets.

Ils sont répartis dans quatre groupes (positif, neutre triste et agressif). Cette variable indépendante est une variable nominale à 4 modalités. Nous avons une relation d'emboîtement entre ce facteur et le facteur sujet.

Le facteur étudié (la variable indépendante) est le type de problème (bien défini ou mal défini), variable nominale dichotomique. Les sujets ne voyant qu'un des deux types de problème, il y a une relation d'emboîtement. La variable dépendante est la performance dans la résolution du problème cible. Aucune précision n'est donnée sur la nature de cette mesure.

Le protocole est donc un protocole univarié structuré. L'objectif de cette recherche est de comparer des groupes d'observations. Le tableau représentant le protocole serait alors le suivant :

Individus	*Humeur*	*Type de problème*	*Performance*
Sujet 1	Positive	Bien défini	VD
Sujet 2	Neutre	Bien défini	VD
Sujet 3	Triste	Mal défini	VD
Sujet 4 …	Agressive …	Mal défini	VD

Exercice 4.

Pour passer des temps exprimés en soixantième de secondes à des temps exprimés en milliseconde, il faut appliquer à chaque modalité de la variable la transformation suivante: x/60*1000.

20 soixantièmes de secondes correspondent donc à 20/60*1000=333 ms (les temps sont ici arrondis pour plus de simplicité).

Exercice 5.

Détermination du nombre de classes souhaité et de l'intervalle de chaque classe:

Nous avons 40 observations, il nous faut donc au maximum 40/5=8 classes. On préférera un nombre impair de classe. L'intervalle de variation étant de 55-20=35 modalités, on optera pour 7 classes parce qu'on a ainsi un nombre entier de modalités dans chaque classe (35/7=5 modalités dans chaque classe) et l'on évite d'étendre l'intervalle de variation.

Détermination de la valeur centrale de la classe centrale.

((55-20)/2)+20=37,5

On arrondi à 38

Détermination des valeurs appartenant à chaque classe.

Nos classes contiennent 5 modalités. Pour la classe centrale, on a donc

38-2=36 et 38+2=40. La classe centrale va donc de 36 à 40.

Pour la classe 3 : 36-5=31 Cette classe va donc de 31 à 35. On procède ainsi pour toutes les autres classes.

On calcule ensuite les valeurs centrales de chacune des classes en faisant la moyenne des valeurs de chaque classe.
On calcule ensuite les limites de classes (somme des valeurs centrales divisée par deux).
Le résultat de ce recodage est synthétisé dans le tableau suivant.

Mini	*Max*	*Val. Centr.*	*Limites*
21	25	23	25,5
26	30	28	30,5
31	35	33	35,5
36	40	38	40,5
41	45	43	45,5
46	50	48	50,5
51	55	53	

Exercice 6.

Concrètement, faire une distribution se fait en deux temps :

1) Lister l'ensemble des modalités observables. Si l'ensemble des modalités observables n'est pas un ensemble fini, on se contentera de lister toutes les modalités entre la plus petite et la plus grande modalité observée.

2) Une fois cette liste établie, on compte pour chaque modalité le nombre de fois où elle apparaît dans le protocole. La distribution du précédent protocole est la suivante :

Modalité u	*Effectif n(u)*	*Modalité u*	*Effectif n(u)*
8	1	30	2
9	1	31	2
10	1	32	3
11	1	33	5
12	1	34	3
13	1	35	4
14	0	36	7
15	0	37	6
16	1	38	3
17	0	39	3
18	3	40	4
19	2	41	7
20	3	42	3
21	2	43	5
22	1	44	7
23	1	45	2
24	4	46	1
25	5	47	1
26	3	48	2
27	3	49	2
28	2	50	1
29	4		

Dans ce protocole, nous avons des observations entre 8 et 50, les modalités entre ces deux valeurs ont donc été listées, y compris celles qui n'ont pas été observées. Pour chacune de ces modalités nous avons compté le nombre de fois où elle a été observée.

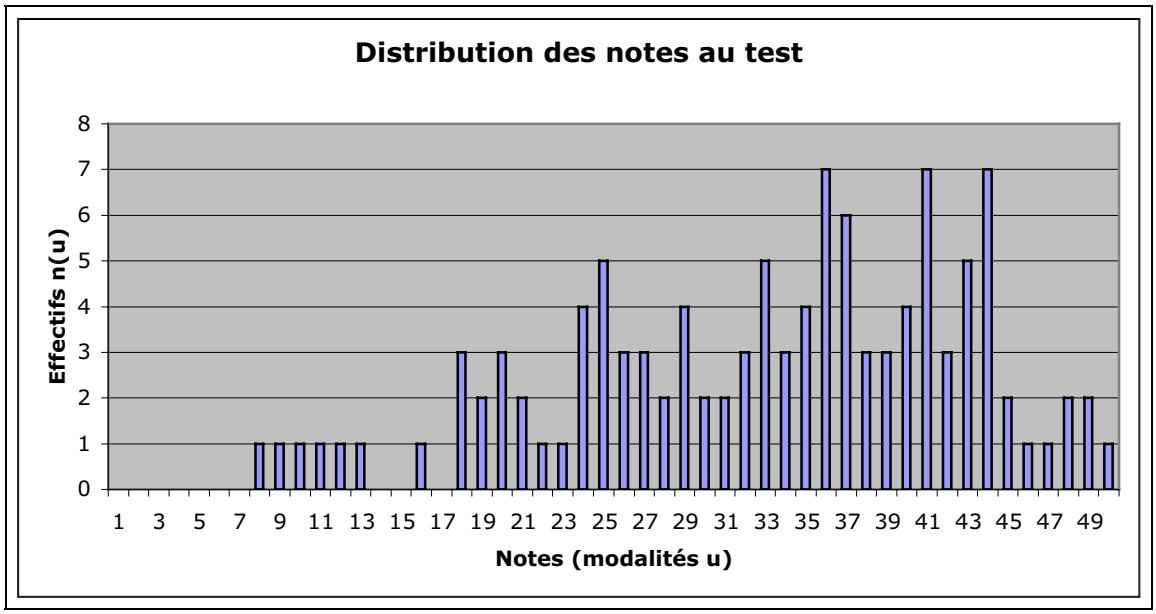
Exercice 7.

On peut représenter une distribution sur les effectifs ou sur les fréquences. Pour cela, on construit un graphique où l'axe des abscisses représente les modalités de la variable et l'axe des ordonnées représente l'échelle des effectifs ou des fréquences. La distribution est représentée par des traits verticaux (bâtonnets) dont la hauteur est proportionnelle à l'effectif ou la fréquence de chaque modalité. Pour des raisons d'esthétique, certains auteurs élargissent ces bâtonnets. Dans ce cas, il faut que tous les bâtonnets aient la même largeur, puisque l'intervalle des classes est le même pour toutes les modalités, c'est-à-dire 1. Un tel graphique

s'appelle un histogramme. La largeur des barres représente la densité d'effectifs de la classe. Dans le cas où l'histogramme serait construit avant regroupement de modalités, la densité d'effectifs est égale à l'effectif, puisque le nombre de modalités dans la classe constituée par la modalité est de 1.

Dans le cas des variables continues, cela peut avoir du sens de relier les sommets des bâtonnets et de construire ainsi une courbe, mais par dans le cas des variables discontinues, puisqu'il n'y a rien entre les modalités.

On peut aussi représenter les distributions à l'aide de représentations par secteur (aussi appelées camembert) où chaque secteur est proportionnel aux effectifs ou aux fréquences. C'est surtout intéressant si on a peu de modalités, sinon on a un nombre de secteur trop important et cela nuit à la lisibilité du graphique. Dans notre exemple, c'est le cas, on préférera donc l'histogramme.



Ce graphique est l'histogramme représentant la distribution des notes au test. On notera que les modalités non observées sont également représentées, et que chacune des barres à une largeur de 1, puisque nous représentons la distribution des notes avant regroupement. Commentons maintenant ce graphique. On peut voir que la distribution est asymétrique. Les observations sont en effet plutôt concentrées vers les notes hautes puisqu'une majorité des observations se situe au-delà du milieu de l'intervalle de variation. Par ailleurs certaines valeurs sont mieux représentées que les autres (37, 41, 44). Nous avons donc ici une distribution multimodale (plusieurs modes).

Exercice 8.

L'énoncé nous donne comme contrainte de faire 7 classes en terminant par la plus importante valeur observée soit 50. Les valeurs extrêmes observées sont 8 et 50. Les notes à ce test étant des nombres entiers, nous avons 50-8+1 = 43 modalités dans l'intervalle observé. Nous prendrons donc comme intervalle de classe le plus petit multiple de 7 qui soit supérieur à 43 soit 49. Nous aurons donc comme intervalle de classe 49/7= 7.

Puisque nous devons terminer par 50, nous commencerons par la dernière classe. A la valeur de la classe la plus élevée, nous retranchons 7-1=6. La dernière classe va donc de 50 à : 50-6= 44

La classe précédente se termine à 43 et commence à 43-6=37. Une autre façon de faire est de retrancher l'intervalle aux valeurs extrêmes de la classe précédente. Pour passer de la dernière classe à l'avant-dernière on aura donc : pour la valeur supérieure : 50-7=43 et pour la valeur inférieure, on aura 44-7=37. L'avant-dernière classe va donc de 37 à 43. On procède ainsi pour toutes les autres classes.

Les limites de classes sont obtenues en additionnant la valeur la plus élevée de la classe et la valeur la plus basse de la classe suivante. On divise ensuite le total par deux. Pour la classe 2-8 on a donc :

Valeur supérieure de la classe : 8

Valeur inférieure de la classe suivante : 9

Limite de la classe 2-8 : (8+9)/2=8,5

On fait ensuite la distribution sur les classes ainsi formées.

Intervalle	*Valeur centrale*	*Limites*	*Effectifs*
2-8	5	8,5	1
9-15	12	15,5	5
16-22	19	22,5	12
23-29	26	29,5	22
30-36	33	36,5	26
37-43	40	43,5	31
44-50	47	50,5	16
		Total	113

Exercice 9.

Nous avons à faire un regroupement en 9 classes sur une échelle comptant 50 modalités. L'intervalle de classe est donc de 50/9= 5,56. Nous prendrons donc un intervalle de 5 modalités. Pour le reste la procédure est la même. Le résultat de ce recodage est le suivant :

Intervalle	*Valeur centrale*	*Limites*	*Effectifs*
6-10	8		3
11-15	13	10,5	3
16-20	18	15,5	9
21-25	23	20,5	13
26-30	28	25,5	14
31-35	33	30,5	17
36-40	38	35,5	23
41-45	43	40,5	24
46-50	48	45,5	7
		Total	113

Exercice 10.

Nous reproduisons ici cette distribution. Le mode correspond à la modalité dont l'effectif est le plus élevé. Dans notre exemple Il s'agit de la classe 41-45. L'effectif de la classe 36-40 est très proche. Cette classe constitue le mode secondaire.

Intervalle	*Valeur centrale*	*Limites*	*Effectifs*
6-10	8	5,5	3
11-15	13	10,5	3
16-20	18	15,5	9
21-25	23	20,5	13
26-30	28	25,5	14
31-35	33	30,5	17
36-40	38	35,5	23
41-45	43	40,5	24
46-50	48	45,5	7
		Total	113

Exercice 11.

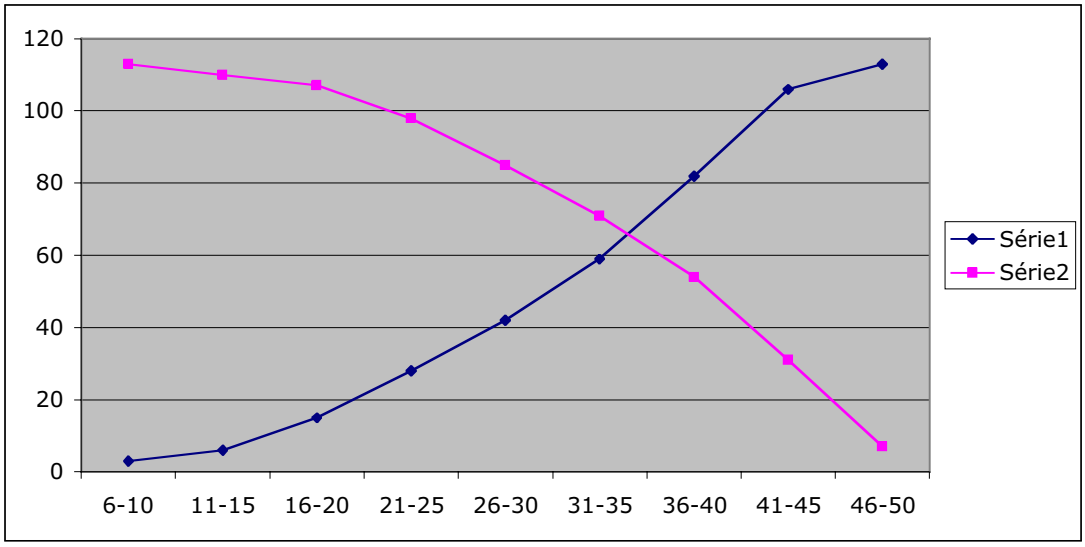
Détermination de la médiane

Pour déterminer la médiane et les quartiles, il faut une distribution cumulée à gauche ou à droite ce qui implique que la variable soit ordinale ou numérique. Nous reprendrons la même distribution en classes pour exemplifier la détermination de la médiane.

min	*max*	*valcentr*	*Effectif*	*effectif cum G*	*effectif cum D*
6	10	8	3	3	113
11	15	13	3	6	110
16	20	18	9	15	107
21	25	23	13	28	98
26	30	28	14	42	85
31	35	33	17	59	71
36	40	38	23	82	54
41	45	43	24	106	31
46	50	48	7	113	7

1) Calculer l'effectif cumulé correspondant à la moitié des observations soit n/2.
Dans notre exemple n vaut 113 et n/2=113/2=66,5
2) Chercher dans les effectifs cumulés n/2. Dans notre cas, on cherchera 66,5. Si une modalité correspond à cet effectif cumulé, on la prend comme médiane. Mais, dans notre exemple, et c'est souvent le cas, cet effectif cumulé ne correspond pas à une modalité. On prendra donc l'effectif cumulé le plus proche. Il s'agit de la classe 31-35. Dans le cas où l'effectif cumulé correspondant à la médiane est à égale distance entre deux modalités, on s'abstiendra de choisir et on parlera de coupure quasi-médiane entre les deux modalités en question. Si plusieurs modalités correspondent à l'effectif cumulé (cas où l'on a des effectifs nuls dans des classes successives), on prend la première modalité.

Détermination graphique de la médiane.



On peut déterminer graphiquement la médiane en faisant sur un même graphique la courbe des effectifs cumulés à gauche (série 1) et à droite (série 2). Pour trouver la médiane, on abaisse la perpendiculaire à l'axe des abscisses qui passe par l'intersection des deux courbes.

Détermination des quartiles.

La procédure est similaire à celle de la détermination de la médiane, mais, dans ce cas, on cherche pour le premier quartile la modalité dont l'effectif cumulé correspond à un quart des observations soit n/4. Pour le troisième quartile, on cherche la modalité dont l'effectif cumulé est égal aux trois quarts des observations soit n*3/4 (le deuxième quartile correspond à la médiane).

Dans notre exemple on cherche donc les modalités dont l'effectif cumulé correspond à :

- Pour le premier quartile 113/4 soit 28,25,
- Pour le troisième quartile 113*3/4=84,75.

On peut voir que la modalité dont l'effectif cumulé est le plus proche de 28,25 est la classe 21-25, elle constituera donc notre premier quartile (Q1). La modalité dont l'effectif cumulé est le plus proche de 84,75 est la classe 36-40. Ce sera notre troisième quartile (Q3).

Interprétation des quartiles et de la médiane.

La médiane nous indique quelle modalité coupe en deux la distribution. Elle montre donc que la moitié des sujets est située en dessous de cet échelon et la moitié au-dessus. Dans notre exemple, nous pouvons donc dire qu'un sujet sur deux appartient aux classes inférieures (ou supérieures ce qui est équivalent) à la classe 31-35.

Un commentaire similaire peut-être fait pour Q1 et Q3, mais, dans ce cas les classes supérieures et inférieures ne sont plus symétriques. On aura ainsi

- Avec Q1 : Un quart des sujets appartient à une classe inférieure à la classe 21-25 et par complémentarité, les trois quarts des sujets, appartiennent à une classe supérieure à la classe 21-25.
- Avec Q3 : Les trois quarts des sujets appartiennent à une classe inférieure à la classe 36-40 et par complémentarité, un quart des sujets appartient à une classe supérieure à la classe 36-40.

Mais, plutôt que de commenter chacun de ces indices séparément, il est plus informatif et plus synthétique de fonder son commentaire sur la combinaison des trois indices. On pourra ainsi dire que la moitié des sujets a obtenu une note comprise entre 23 (valeur centrale de Q1) et 38 (valeur centrale de Q3) avec une médiane à 33 (valeur centrale de la classe médiane). Puisqu'on sait, avec la distribution, que les notes observées vont de 8 à 50, on situe tout de suite la répartition des observations du côté des valeurs élevées, mais aussi que la répartition des notes autour de la médiane n'est pas symétrique.

Exercice 12.

Commençons par la moyenne. On sait calculer une moyenne depuis l'école primaire, Il s'agit de la somme des observations divisée par le nombre d'observations. Mais lorsqu'on a un ensemble de données important, cela peut créer quelques difficultés. Nous noterons la somme des observations Σx, le total des observations n et la moyenne m.

On a donc $m = \Sigma x/n$.

En pratique, pour calculer une moyenne, on peut partir du protocole ou de la distribution. Il est tout à fait important de bien différencier les deux, car la procédure n'est pas la même. Il est

rare que le protocole se présente sous la forme d'une série de données en vrac. On les présente en général sous la forme d'un tableau et c'est ce qui induit en erreur bon nombre d'étudiants parce qu'ils ne savent pas différencier le tableau d'un protocole de celui d'une distribution. Pour bien les différencier, il convient de se demander à quoi correspondent chacune des lignes. Dans un tableau de protocole les lignes correspondent aux individus statistiques, dans une distribution, les lignes correspondent aux modalités de la variable. Bien rangé, le tableau de protocole correspondant à notre exemple aurait l'allure suivante :

individus (i)	*Observation (x)*
1	43
2	31
3	38
...	...
110	37
111	48
112	29
113	35

Il fait correspondre à un ensemble d'individus (notés i) un ensemble d'observations (notées x). Chaque observation relative à un sujet est notée x_i (le i renvoyant à l'individu statistique). La somme des observations se notera donc Σx_i. La moyenne sera $m=\Sigma x_i/n$. Concrètement cela veut dire qu'on fait la somme de la colonne des observations et qu'on la divise par le nombre des observations. Dans notre exemple, on aura :

(43+31+38+...+37+48+29+35)/113=32,95

Calcul de la moyenne à partir de la distribution.

Si on part de la distribution, la procédure est un peu différente. Le tableau de distribution nous donne le nombre de fois où chaque modalité a été observée (effectif). Il convient alors, pour faire le total des observations, de multiplier les modalités par leur effectif. Pour bien comprendre cela, imaginez que dans votre bulletin scolaire vous ayez obtenu 3 fois la note 9 et 2 fois la note 11 (oui, je sais, ce n'est pas cher payé), vous comprenez aisément qu'il est équivalent pour calculer la moyenne de faire :

(9+9+9+11+11)/5=9,8 (ce qu'on ferait en partant du protocole)

ou

(9*3+11*2)/5=9,8 (ce qu'on fait en partant de la distribution)

Concrètement, je viens de résumer le protocole (votre bulletin scolaire) sous la forme d'une distribution qu'on pourrait représenter sous la forme du petit tableau suivant :

Notes (ou modalités de la variable)	*Effectifs (ou nombre de notes)*
9	3
11	2

Si vous partez de la distribution, il vous faudra
1) Multiplier chaque modalité par son effectif,
2) Faire la somme de ces produits,
3) Diviser par le total des effectifs (nombre d'observations).

Un peu de formalisme.
Appelons maintenant u^k les modalités de la variable et n^k les effectifs (pour faire simple, nous interpréterons k comme un renvoi à une ligne du tableau de distribution sans tenir compte de sa position en indice ou en exposant)
La somme des observations est alors $\Sigma n_k u^k$, ce qui se lit somme des produits de n_k par u^k.
Et la moyenne est alors égale à $\Sigma n^k u_k$ /n.
Concrètement sur l'exemple des notes au test, on construira le tableau suivant :

u^k	n_k	$n_k u^k$
6	0	0
7	0	0
8	1	8
...	...	...
...	...	...
47	1	47
48	2	96
49	2	98
50	1	50
Total	113	3723

Et la moyenne de 3723/113=32,95

Calcul de la variance et de l'écart-type
En pratique, on ne peut pas calculer directement l'écart-type. Il faut pour cela d'abord calculer la variance et ensuite extraire sa racine carrée. Comme pour la moyenne (mais après tout la variance est aussi une moyenne), on peut partir du protocole ou de la distribution. Dans ces deux cas, on a deux formules permettant de faire ce calcul: la formule de définition et la formule dite "de calcul" parce qu'elle est plus souvent utilisée car elle permet d'éviter les erreurs liées aux arrondis.
Rappelons ici la définition de la variance (il va falloir s'en imprégner) :

La variance est la moyenne des carrés des écarts à la moyenne.

On note la variance : s^2 et l'écart-type : s. On trouvera aussi la notation suivante pour la variance : var ou var(x).

Concrètement qu'est-ce que cela signifie ?

Les écarts à la moyenne s'écrivent (x_i-m) ce qui signifie qu'il faut soustraire la moyenne du protocole à chaque observation.

Le carré des écarts à la moyenne s'écrira donc $(x_i-m)^2$, on élève donc chacun de ces écarts au carré.

La moyenne de ces carrés des écarts à la moyenne s'écriera donc $s^2 = \Sigma(x_i-m)^2/n$. Ce qui signifie qu'on va faire la somme des carrés des écarts à la moyenne et les diviser par le nombre d'observations.

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n}$$

Utilisation de la formule de définition à partir du protocole.

Cette petite explication détaillée de la formule de définition vous donne la procédure pour calculer la variance à partir du protocole. Concrètement voici comment cela se passe.

On reprend le tableau du protocole, puis on ajoute deux colonnes, l'une pour calculer les écarts à la moyenne, l'autre pour calculer les carrés de ces écarts.

On calcule la moyenne.

- Pour chaque individu statistique, on calcule son écart à la moyenne soit (x_i-m)
  Pour i_1, on a : 43-32,95=10,05
  Pour i_2, on a : 31-32,95=-1,95
  Pour i_3, on a : 38-32,95=5,05
- Pour chaque individu statistique, on élève son écart à la moyenne au carré soit $(x_i-m)^2$
  Pour i_1, on a : 10,052=101,06
  Pour i_2, on a : -1,952=3,79
  Pour i_3, on a : 5,052=25,53
- On fait ensuite la somme de ces carrés qu'on divise par le nombre d'individus statistiques soit $\Sigma(x_i-m)^2 /n$

s^2 = (101,06+3,79+25,53+...+15,58+4,22) / 113 = 97,06

m=32,95

i	*xi*	*xi-m*	*(xi-m)2*
1	43	10,05	101,06
2	31	-1,95	3,79
3	38	5,05	25,53
...	...		
110	37	4,05	16,43
111	48	15,05	226,6
112	29	-3,95	15,58
113	35	2,05	4,22
total	3723	0	10967,68

L'écart-type est la racine carrée de la variance soit : $s = \sqrt{\text{var}} = \sqrt{97,06} = 9,85$

Utilisation de la formule de définition à partir de la distribution

La difficulté ici est similaire à celle rencontrée lors du calcul de la moyenne à partir de la distribution. On peut bien calculer l'écart à la moyenne de chaque modalité, mais il ne faut pas oublier de multiplier cet écart par le nombre de fois où il a été observé, c'est-à-dire par l'effectif.

Concrètement, on commence par préparer le tableau de distribution en lui ajoutant trois colonnes:
Une pour le calcul de l'écart à la moyenne.
Une autre pour l'élévation de cet écart au carré
et une troisième pour le produit de ce carré par l'effectif de la modalité.

On calcule l'écart de chaque modalité avec la moyenne du protocole soit $u^k - m$.

Pour u^1, on a : 32,95-6 = -26,95
Pour u^2, on a : 32,95-7 = -25,95
Pour u^3, on a : 32,95-8 = -24,95 etc.

Pour chaque modalité, on calcule le carré de cet écart, soit $(u^k - m)^2$.

Pour u^1, on a : $-26,95^2 = 726,14$
Pour u^2, on a : $-25,95^2 = 673,24$
Pour u^3, on a : $-24,95^2 = 622,35$ etc.

Pour chaque modalité, on calcule le produit de cet écart et de son effectif soit: $(u^k - m)^2 * n_k$

Pour u^1, on a :726,14 * 0 = 0
Pour u^2, on a : 673,24* 0 = 0
Pour u^3, on a : 622,35 * 1 = 622,35 etc.

On fait ensuite la somme de ces produits et on la divise par n soit $s^2=\Sigma((u^k - m)^2*n_k) / n$.
$s^2=(0+0+622,35+...+515,40+290,81) / 113 = 10967,68/113 = 97,06$

u^k	n_k	u^k-m	$(u^k$-m)2	$n(u^k$-m)^2
6	0	-26,95	726,14	0
7	0	-25,95	673,24	0
8	1	-24,95	622,35	622,35
46	1	13,05	170,38	170,38
47	1	14,05	197,49	197,49
48	2	15,05	226,6	453,19
49	2	16,05	257,7	515,4
50	1	17,05	290,81	290,81
Total	113		Total	10967,68

L'écart-type se calcule de la même façon que précédemment variance soit $s = \sqrt{\text{var}} = \sqrt{97,06} = 9,85$

Utilisation de la formule de calcul

On peut montrer que la formule de définition $\sum (x_i - m)^2 / n$ est équivalente à la formule suivante:

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n}$$

Cette deuxième formule signifie qu'il faut faire la différence entre la somme des carrés des observations (Σx^2) et le carré de la somme des observations pondérée par n soit $(\Sigma x)^2/n$) et diviser ensuite cette différence par n.
Pour ceux que cela intéresse, essayez de retrouver la suite de transformations algébriques qui permettent de passer de l'une à l'autre. Pour les autres, passons à sa mise en application. Encore une fois, nous devons différencier l'utilisation du protocole ou de la distribution comme point de départ des calculs.

Utilisation de la formule de calcul à partir du protocole.
Concrètement, il va falloir calculer la somme des carrés et la somme des observations pour pouvoir appliquer la formule. On prépare donc un tableau correspondant à son protocole et on lui ajoute une colonne pour calculer les carrés des observations. On voit tout de suite l'avantage de cette procédure, Puisque la somme des observations a déjà été calculée pour la moyenne, il n'y a plus qu'une colonne de calculs à faire.

Pour chaque observation, on calcule son carré.

- Pour i_1, on a : 43^2=1849
- Pour i_2, on a : 31^2=961
- Pour i_3, on a : 38^2=1444

On fait ensuite la somme des carrés.

Dans notre exemple, Σx^2= 133629

Si ce n'est déjà fait pour la moyenne, on calcule la somme des observations.

Dans notre exemple, Σx = 3723

i	x_i	x_i^2
1	43	1849
2	31	961
3	38	1444
4	40	1600
...	...	....
110	32	1024
111	36	1296
112	44	1936
113	37	1369
Total	3723	133629

Dans la mesure où l'application de la formule est la même que l'on parte d'un protocole ou d'une distribution, nous développerons son application à la fin.

Utilisation de la formule de calcul à partir de la distribution.

Le problème est toujours le même dans le passage du protocole à la distribution pour le calcul de ces indices, il ne faut pas oublier de multiplier par l'effectif. Concrètement on prépare un tableau de distribution en lui ajoutant deux colonnes. La première servira à calculer le carré des modalités soit u^{k2}, la seconde servira à multiplier ce carré par l'effectif de la modalité soit $n_k(u^{k2})$. Si la somme des observations n'a pas été calculée pour la moyenne, on aura besoin d'une troisième colonne pour calculer le produit de chaque modalité par son effectif soit $n_k u^k$.

On calcule alors la somme des carrés des observations (Σx^2)

Pour chaque modalité, on calcule son carré. Par exemple,

- Pour u^1, on a 6^2=36
- Pour u^2, on a 7^2=49
- Pour u^3, on a 8^2=64 etc.

Pour chaque modalité, on multiplie le carré par l'effectif de la modalité.

- Pour u^1, on a 36*0=0
- Pour u^2, on a 49*0=0
- Pour u^3, on a 64*1=64 etc.

Attention, il s'agit bien ici de multiplier le carré de la modalité par son effectif et non de faire le carrré du produit de l'effectifs par la modalité. Ainsi, pour la note 48, on a bien : $48^2*2=2401*2=4802$.
Et non de faire le carré du produit de l'effectif par la modalité.
$(48*2)^2=9216$

On fait ensuite la somme de ces produits. Dans notre exemple, $\Sigma x^2=133629$

Si ce n'est déjà fait lors du calcul de la moyenne, on calcule la somme des observations (Σx)
On multiplie chaque modalité par son effectif, par exemple

- Pour u^1, on a 6*0=0
- Pour u^2, on a 7*0=0
- Pour u^3, on a 8*1=8 etc.

On fait la somme de ces produits. Dans notre exemple, $\Sigma x=3723$

u^k	n_k	u^{k2}	$u^{k2}*n$	$u^k n_k$
6	0	36	0	0
7	0	49	0	0
8	1	64	64	8
9	1	81	81	9
...	...	...	...	...
47	1	2209	2209	47
48	2	2304	4608	96
49	2	2401	4802	98
50	1	2500	2500	50
Total	113		133629	3723

On applique ensuite la formule en calculant dans l'ordre

- Le carré de la somme des observations $(\Sigma x)^2 = 3723^2=13860729$
- On divise ensuite ce carré par n, soit $(\Sigma x)^2/n =13860729/113=122661,319$
- On fait ensuite la différence avec la somme des carrés, soit :

$$\Sigma x^2-((\Sigma x)^2/n=133629-122661,319=10967,68$$

Enfin on divise le tout par n soit :

$$S^2=(\Sigma x^2-((\Sigma x)^2/n)/n = 10967,68/113=97,06$$

L'écart-type se calcule de la même façon que précédemment. Il est égal à la racine carrée de la variance, soit s=9,85.

Le résultat est bien sûr identique à celui que nous avions trouvé précédemment. En pratique, ce n'est pas toujours le cas, notamment si on doit effectuer le calcul à la main. Cela tient au fait que dans l'utilisation de la formule de définition, on est amené à faire beaucoup plus d'arrondis (dans les différences avec la moyenne) ce qui entraîne une perte de précision. **Il est donc préférable d'utiliser la formule de calcul.**

Exercice 13.

Pour faire un décilage, il est nécessaire d'avoir suffisamment de modalités (au moins 20 ou 30) sinon les déciles vont se chevaucher et le résultat ne sera pas très informatif. L'échelle de mesure de la variable doit être ordinale ou numérique et on doit disposer d'une distribution cumulée (à gauche ou à droite). Il est plus commode de partir d'une distribution cumulée des fréquences plutôt que des effectifs (pour ne pas avoir à calculer à chaque fois l'effectif cumulé correspondant). Les fréquences cumulées à rechercher dans la distribution cumulée des fréquences pour le protocole sont les suivantes.

N° du décile	*Fréquence cumulée à gauche*	*Fréquence cumulée à droite*
1	0,1	0,9
2	0,2	0,8
3	0,3	0,7
4	0,4	0,6
5	0,5	0,5
6	0,6	0,4
7	0,7	0,3
8	0,8	0,2
9	0,9	0,1

La procédure consiste à relever la modalité correspondant à chacune de ces fréquences cumulées dans le tableau des fréquences cumulées du protocole.Si une modalité correspond à cette fréquence cumulée, on la prend pour coupure, sinon on cherche la modalité la plus proche avant ou après. Si la fréquence du décile est à mi chemin entre deux modalités, on s'abstient de choisir et on prend l'intervalle comme coupure interdécile. Dans le cas où plusieurs modalités correspondent à la fréquence du décile (effectif nul dans plusieurs classes successives), on prend la première modalité correspondant à la fréquence cherchée.

Modalité	*Effectif*	*Fréq*	*Fréq cum.*	*Coupure*	*Modalité*	*Effectif*	*Fréq*	*Fréq cum.*	*Coupure*
8	1	0,01	0,01		32	3	0,03	0,42	
9	1	0,01	0,02		33	5	0,04	0,46	
10	1	0,01	0,03		34	3	0,03	0,49	34
11	1	0,01	0,04		35	4	0,04	0,52	
12	1	0,01	0,04		36	7	0,06	0,58	36
13	1	0,01	0,05		37	6	0,05	0,64	
14	0	0	0,05		38	3	0,03	0,66	
15	0	0	0,05		39	3	0,03	0,69	39
16	1	0,01	0,06		40	4	0,04	0,73	
17	0	0	0,06		41	7	0,06	0,79	
18	3	0,03	0,09	18/19	42	3	0,03	0,81	41/42
19	2	0,02	0,11		43	5	0,04	0,86	
20	3	0,03	0,13		44	7	0,06	0,92	44
21	2	0,02	0,15		45	2	0,02	0,94	
22	1	0,01	0,16		46	1	0,01	0,95	
23	1	0,01	0,17		47	1	0,01	0,96	
24	4	0,04	0,2	24	48	2	0,02	0,97	
25	5	0,04	0,25		49	2	0,02	0,99	
26	3	0,03	0,27		50	1	0,01	1	
27	3	0,03	0,3	27					
28	2	0,02	0,32						
29	4	0,04	0,35						
30	2	0,02	0,37						
31	2	0,02	0,39	31					

À partir des déciles, on définit des classes (comprise entre deux déciles) appelées interdéciles. Dans notre exemple, les interdéciles sont les suivants :

Interdécile N°1	8-18
Interdécile N°2	19-24
Interdécile N°3	25-27
Interdécile N°4	28-31
Interdécile N°5	32-34
Interdécile N°6	35-36
Interdécile N°7	37-39
Interdécile N°8	40-42
Interdécile N°9	43-44
Interdécile N°10	45-50

Voyons maintenant comment situer dans la distribution un sujet qui aurait eu une note de 25. note est plutôt faible, puisque le sujet appartient au 3ème interdécile. Ce qui veut dire que 70 % des sujets ont une note supérieure. Notre sujet a pourtant la moitié des points, mais sur ce test, ce n'est pas beaucoup. La construction des interdéciles correspond à ce qu'on appelle un étalonnage en déciles. Elle est très utilisée dans la construction des tests. Elle permet ainsi de

situer la performance d'un sujet par rapport aux résultats obtenus dans un échantillon de référence (voir le cours sur les méthodes et notamment la méthode des tests).

Exercice 14.
Ce calcul ne pose pas de problème particulier, Il s'agit d'une simple soustraction et une division. Voyons ce que cela donne pour notre sujet qui a eu 25 au test. Rappelons que la moyenne à ce test est de 32,95 et l'écart-type de 9,85.

$$z= x_i-m / s = 25 - 32,95 / 9,85 = -0,81$$

On voit que la note z est négative. La distance à la moyenne est donnée par la valeur absolue de la note z. Elle signifie, dans cet exemple, que notre sujet est situé à 0,81 écart-type de la moyenne. Le signe de la note z indique le sens de cet écart. Si la note est négative, le sujet est situé en dessous de la moyenne. Si le signe est positif, le sujet est situé au-dessus de la moyenne. Dans notre exemple, il est en dessous. Concrètement qu'est-ce que cela veut dire ?
La note z exprime l'écart à la moyenne en nombre d'écart-type. Cette note z de -0,81 veut dire que notre sujet est à moins d'un écart-type de la moyenne. Autrement dit, il ne diffère pas beaucoup de la tendance générale de l'échantillon, même si il est du côté des valeurs faibles.

Exercice 15.
Nous avons utilisé deux moyens de situer notre sujet dans la distribution et ils nous apportent des réponses un peu contradictoires. Notre sujet serait moyen et pourtant 70 % des sujets ont une note supérieure. Pourquoi ? C'est que la moyenne et l'écart-type ne reflètent pas la répartition des observations dans la distribution, mais seulement son centre de gravité et la dispersion autour de ce centre. Sauf dans un cas, celui où l'on a une distribution normale. La contradiction entre nos deux méthodes résulte du fait que dans cet exemple, la distribution n'est pas une distribution normale. C'est pourquoi nous allons la normaliser. Voici la procédure :

Le point de départ est une distribution cumulée. Il vaut mieux le faire avec une distribution cumulée des fréquences pour ne pas avoir à calculer l'effectif cumulé correspondant à chaque coupure. Mais à titre d'exemple, nous montrerons comment faire à partir des effectifs cumulés. Concrètement, ce que nous allons faire est un recodage de la variable par regroupement de modalités. Comme précédemment, il est plus judicieux de choisir un nombre impair de classes. En pratique on choisit une dizaine de classes. Les nombres les plus proches de 10 sont 9 et 11, ce sont donc ceux qu'on utilise le plus, mais rien n'interdit, en fonction de la finesse de l'échelle souhaitée d'en choisir d'autres. Dans une distribution normale, 95,5 % des observations sont à moins de deux écarts-types de la moyenne. C'est donc entre ces deux bornes (m-2s et m+2s) qu'on situera notre distribution. L'intervalle de variation qu'on souhaite obtenir est donc de 4.

i) Calculer l'intervalle de classe en note z

Comme précédemment avec les quartiles et les déciles, le nombre de coupures nécessaires est égal au nombre d'intervalles à obtenir moins un.

Pour obtenir 11 classes, il nous faut 10 coupures. L'intervalle de classe est donc de 4/10=0,4.

Pour obtenir 9 classes, il nous faut 8 coupures. L'intervalle de classe est donc de 4/8=0,5.

Pour obtenir 7 classes, il nous faut 6 coupures. L'intervalle de classe est donc de 4/6=0,67 (valeur arrondie).

Pour notre exemple des notes au test, nous avons choisi de regrouper les modalités en 11 classes.

ii) Calculer les limites de la classe centrale en note z.

Puisque la distribution z est centrée sur 0, la classe centrale doit également être centrée sur 0. Les limites de cette classe sont égales à un demi-intervalle en plus ou en moins autour de 0. Pour 11 classes, ces limites seront 0,4/2=0,2 donc +0,2 et -0,2.

iii) Calculer les limites des autres classes.

L'intervalle entre deux limites de classes est égal à l'intervalle de classe. Pour trouver les autres limites de classe, il suffit en partant des classes centrales de retrancher ou d'additionner cet intervalle (dans notre exemple, l'intervalle de classe est de 0,40).

Du côté des valeurs négatives, on retranche un intervalle de classe à la limite de la classe supérieure. Nous aurons donc -0,20 ; -0,60 ; -1 ; -1,4 ; -1,8

Du côté des valeurs positives, on retranche un intervalle de classe à la limite de la classe inférieure. nous aurons 0,20 ; 0,60 ; 1 ; 1,4 ; 1,8.

Ces limites de classes sont bien sûr symétriques puisque nous cherchons une distribution centrée sur 0.

iv) Consultation de la table de la loi normale.

On consulte ensuite la table de distribution cumulée à gauche de la loi normale réduite (appelée aussi "table de z"). Cette table nous donne, pour chaque valeur de z (appelée u dans la table; rappelons que la lettre u désigne les modalités de la variable), la fréquence cumulée à gauche de ces notes (p(z<u) , ce qui se lit proportion de notes z inférieures à u) dans une distribution normale. On lira dans la table la proportion associée à chacune des limites de classe. Par exemple :

- ✓ Pour la limite de classe -1,8 on peut lire dans la table 0,036.
- ✓ Pour la limite de classe -1,4 on peut lire dans la table 0,081 etc.

v) Calcul des effectifs cumulés correspondants à ces fréquences

Pour chacune des classes, on calcule l'effectif cumulé correspondant en multipliant la fréquence (p(z<u)) par l'effectif total n. comme les effectifs sont des nombres entiers, on arrondi à l'entier supérieur ou inférieur le plus proche. Dans notre exemple n=113 on a donc :

- ✓ Pour la coupure 1 : 0,036*113 = 4,068 soit environ 4.
- ✓ Pour la coupure 2 : 0,081*113 = 9,153 soit environ 9.
- ✓ Pour la coupure 3 : 0,159*113 = 17,97 soit environ 18.

vi) Détermination des coupures (limites de classes) en notes au test.

La procédure de détermination des coupures est similaire à celles des déciles. Pour chacune des coupures, on cherche, dans la distribution cumulée des notes au test, la modalité dont l'effectif cumulé est le plus proche de l'effectif cumulé qu'on vient de calculer (qn). La procédure consiste à relever les modalités correspondant à chacun de ces effectifs cumulées dans le tableau des effectifs cumulés du protocole.

Si une modalité correspond à cet effectif cumulé, on la prend pour coupure (nous l'appellerons n' et la modalité suivante sera n").

- Sinon on cherche la modalité la plus proche avant (n') ou après (n").
- Si l'effectif cumulé est à mi chemin entre deux modalités, on s'abstient de choisir et on prend comme coupure n'+n"/2.

Dans le cas où plusieurs modalités correspondraient à l'effectif cumulé (effectif nul dans plusieurs classes successives), on prend la première modalité correspondant à l'effectif cherché.

Modalité	*Effectif*	*effectif cumulé*	*Coupure*	*Modalité*	*Effectif*	*effectif cumulé*	*Coupure*
6	0	0		33	5	52	
7	0	0		34	3	55	
8	1	1		35	4	59	
9	1	2		36	7	66	6
10	1	3		37	6	72	
11	1	4	1	38	3	75	
12	1	5		39	3	78	
13	1	6		40	4	82	7
14	0	6		41	7	89	
15	0	6		42	3	92	
16	1	7		43	5	97	8
17	0	7		44	7	104	9
18	3	10	2	45	2	106	
19	2	12		46	1	107	
20	3	15		47	1	108	
21	2	17		48	2	110	10
22	1	18	3	49	2	112	
23	1	19		50	1	113	
24	4	23					
25	5	28					
26	3	31	4				
27	3	34					
28	2	36					
29	4	40					
30	2	42					
31	2	44					
32	3	47	5				

Dans notre exemple, on a :

- ✓ Pour la classe 1 : qn=4 ; cet effectif correspond à celui de la modalité 11, donc n'=11 et n"=12
- ✓ Pour la classe 2 : qn=9; cet effectif correspond à celui de la modalité 18, donc n'=18 et n"=19 etc.
- ✓

vii) Détermination des classes

Une petite lapalissade nous donnera la procédure. Chaque classe commence là où se termine la précédente et se termine là où commence la suivante.

n' est la coupure, elle signale la fin de la classe. n" est la modalité suivante, elle indique le début de la classe suivante.

La première classe va donc de 0 à 11 (n')

La seconde classe va de 12 (n" de la classe précédente) à 18 (n')

La troisième classe va de 19 (n" de la classe précédente) 22 (n') etc.

Une autre procédure consiste à calculer les limites de classe. Elles se calculent de la manière suivante : Limite de classe = (n'+n")/2.

Dans notre exemple, la limite de classe 1/2 est de (11+12)/2=11,5, ce qui revient à dire que tout ce qui est avant 11,5 appartient à la classe 1 (elle va donc de 0 à 11) et tout ce qui est après appartient à la classe 2 (elle commence donc à 12). Cette procédure est plus facile si on travaille sur des échelles de mesure continue avec plusieurs observations entre deux valeurs entières. Les deux procédures sont équivalentes.

viii) On calcule enfin les effectifs correspondant à chacune des classes. Cette distribution est appelée distribution normalisée. Nous la résumons dans le tableau ci-dessous.

N° classe	*Limite (note z)*	*p(z<u)*	*effectif cumulé*	*n'*	*n"*	*limites*	*Classes*	*Effectif*
1	-1,8	0,036	4,068	11	12	11,5	0 à 11	4
2	-1,4	0,081	9,153	18	19	18,5	12 à 18	6
3	-1	0,159	17,967	22	23	22,5	19 à 22	8
4	-0,6	0,274	30,962	26	26	26	23 à26	13
5	-0,2	0,421	47,573	32	32	32	27 à 32	16
6	0,2	0,579	65,427	36	36	36	33 à 36	19
7	0,6	0,726	82,038	40	40	40	37 à 40	16
8	1	0,841	95,033	43	43	43	41 à 43	15
9	1,4	0,919	103,85	44	45	44,5	44 à 44	7
10	1,8	0,964	108,93	48	49	48,5	45 à 48	6
							49 à 50	3
							Total.	113

Revenons à notre sujet dont la note est de 25. Qu'en est-il ? Rappelons que la note z de notre sujet est de :

$$z= xi-m / s = 25 - 32,95 / 9,85 = -0,81$$

Il appartient à la 4ème classe, celle qui comprend les notes de 23 à 26. On peut voir que seulement 27 % des sujets ont une note inférieure à notre sujet (il faut lire la colonne p(z<u) dans le tableau). Bien qu'il ne soit qu'à moins d'un écart-type de la moyenne, nous pouvons donc dire que sa performance n'a pas été très bonne. On remarquera la proximité des conclusions tirées à partir de la normalisation et du décilage, puisque précédemment nous avions conclu que 70 % des sujets avaient une note supérieure à 25. Cependant il n'en va pas ainsi de tous les sujets. Si nous prenons par exemple un sujet qui aurait eu 27, dans le décilage, il aurait appartenu à la même classe qu'un sujet qui aurait eu 25. Alors que dans la distribution normalisée, un tel sujet appartient à la classe 5, pour laquelle nous avons 42 % d'observations inférieures. Cela tient à la fois au nombre de classe (ici nous en avons 11, alors qu'il n'y en a que 10 dans le décilage) et au fait que la répartition des observations dans les classes dépend de l'écart à la moyenne après normalisation.

Exercice 16.

Dans cette petite expérience, nous avons une variable nominale, les individus statistiques sont les sujets. Des réponses au hasard présupposent que chaque réponse ait autant de chance que les autres d'être choisie par les sujets. Ce qui revient à faire l'hypothèse d'une distribution uniforme sur les trois réponses possibles. Nous allons évaluer l'écart entre la distribution observée et une distribution uniforme en calculant la statistique X^2.

La première étape consiste à calculer les effectifs théoriques. Dans le cas d'une distribution uniforme, ils correspondent au nombre d'observations divisées par le nombre de modalités. Dans notre exemple, on aura donc 60/3=20

On calcule ensuite la différence entre les effectifs observés et les effectifs théoriques. Pour la première modalité, on aura donc : 25-20=5

Chacune de ces différences est élevée au carré. Pour la première modalité, on aura donc 5^2=25

On fait ensuite la somme de ces carrés qu'on divise par l'effectif théorique. Ce qui nous fait : X^2=(25+0+25)/20=2,5

Conclusions possibles.	*Effectifs observés*	*Effectifs théoriques*	*Ecarts*	*Carrés des écarts*
Je suis riche.	25	20	5	25
Je ne suis pas riche.	20	20	0	0
On ne peut pas savoir.	15	20	-5	25
Total	60			2,5

Bien que ce ne soit pas présenté dans votre cours, cette utilisation de X^2 pose parfois quelques problèmes. D'abord parce que la statistique X^2 est dépendante de l'effectif total. Vous pourrez vérifier à titre d'exercice que lorsqu'on double l'effectif total, X^2 double. Pour décider si

l'écart entre la distribution observée et la distribution théorique est important, il faut donc le rapporter à l'effectif total.

Le second problème, spécifique à ce cas un peu particulier de la comparaison d'une distribution à une distribution uniforme, c'est que X^2 varie également en fonction du nombre de modalités (pour ceux que cela amuse, regarder ce qui se passe lorsqu'on double le nombre de modalités).

Il est donc nécessaire de pondérer X^2 par l'effectif total. Ce dernier indice est le carré moyen de contingence Φ^2 (Phi deux). Dans notre exemple, on aura donc : Φ^2= 2,5/60= 0,042. Φ^2 varie dans ce cas de 0 à 2, c'est-à-dire le nombre de modalités moins un. On peut observer que le carré moyen de contingence que nous avons calculé est très proche de 0. On pourra donc considérer que la distribution observée est proche d'une distribution uniforme.

Exercice 17.

Le point de départ est une distribution cumulée. Nous choisirons de la cumuler à gauche puisque la table fournie dans votre cours est la table de fonction cumulée à gauche (p594).

- Calcul des notes z correspondant aux limites de classes. Pour cela nous avons besoin de la moyenne et de l'écart-type. Nous pourrions reprendre les résultats de nos précédents calculs sur la distribution avant regroupements, mais à titre d'exercice, vous les calculerez sur la distribution après regroupement.

D-. Puisqu'on part d'une distribution, il nous faudra calculer pour chaque ligne du tableau le produit de la valeur centrale de la classe (u^k) et de son effectif (n_k) On aura ainsi u^1n_1=3*8=24 ; u^2n_2=3*13=39 etc.

E-. On calculera ensuite le produit du carré de la valeur central par son effectif ($u^{k2}n_k$). Pour u^1, on aura ainsi 8^2*3=192 ; pour u^2, 13^2*3=507 etc.

Intervalle	*Valeur centrale*	*Limites*	*Effectifs*	$u^k n_k$	$u^{k2} n_k$
6-10	8		3	24	192
11-15	13	10,5	3	39	507
16-20	18	15,5	9	162	2916
21-25	23	20,5	13	299	6877
26-30	28	25,5	14	392	10976
31-35	33	30,5	17	561	18513
36-40	38	35,5	23	874	33212
41-45	43	40,5	24	1032	44376
46-50	48	45,5	7	336	16128
		Total	113	3719	133697

On a donc comme moyenne : m= 3719/113= 32,91.

La variance est de : $s2 = \frac{\sum u^{k2}n^k - \frac{(\sum u^k n_k)^2}{n}}{n} = \frac{133697 - \frac{(3719)^2}{113}}{113} = 99{,}99$

Et l'écart-type est de : $s = \sqrt{99,99} \approx 10$

Le calcul des notes z pour chaque classe est obtenu en faisant la différence entre limite de classe et la moyenne qu'on divise par l'écart-type. On a ainsi pour la première classe : z=(10,5-32,91)/10=-2,24 ; pour la deuxième classe : z=(15,5-32,91)/10=-1,74 etc.

- Détermination des fréquences cumulées théoriques. Pour chacune des classes, on relève, dans la table de la fonction z, les fréquences théoriques correspondant à la note z de la classe. Dans la première classe, z=-2,24 soit environ -2,2. Dans la table, au regard de cette note z, on lit 0,014. De la même manière, pour la classe 2, on lit pour z=-1,74 la valeur 0,041. On procède ainsi pour toutes les classes.

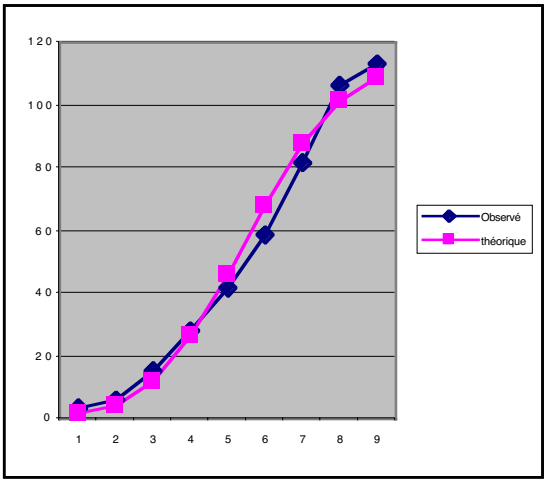
- Calcul des effectifs cumulés théoriques. Il s'obtient en multipliant la note z de la classe par l'effectif total. On a ainsi pour la première classe : 0,014*113=1,582 ; Pour la deuxième classe, on a 0,041*113=4,633 etc.

- Calcul des effectifs théoriques non cumulés. Pour cela, on fait la différence entre l'effectif cumulé de la classe et l'effectif cumulé de la classe précédente.

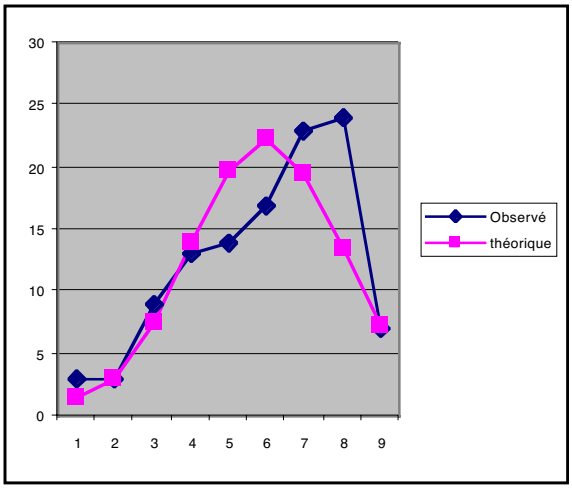
Pour la classe 6-10, on a donc 1,582-0=1,582 (ici l'effectif cumulé précédent est égal à 0 puisqu'il n'y a pas de classe précédente).
Pour la classe 11-15, on a 4,633-1,582=3,501.
Pour la classe 16-20, on a 12,091-4,633=7,458 etc.

Intervalle.	*Limites.*	*Notes Z.*	*Fréquences théoriques cumulées.*	*Effectifs cumulés théoriques.*	*Effectifs théoriques.*
6-10	10,5	-2,24	0,014	1,582	1,582
11-15	15,5	-1,74	0,041	4,633	3,051
16-20	20,5	-1,24	0,107	12,091	7,458
21-25	25,5	-0,74	0,230	25,99	13,899
26-30	30,5	-0,24	0,405	45,765	19,775
31-35	35,5	0,26	0,603	68,139	22,374
36-40	40,5	0,76	0,776	87,688	19,549
41-45	45,5	1,26	0,896	101,248	13,56
46-50	50,5	1,76	0,961	108,593	7,345

- comparaison graphique entre les deux distributions. On représente pour cela sur un même graphique la distribution (cumulée ou non) observée et théorique. Nous donnons ici les représentations graphiques sur les distributions cumulées (graphique de droite) et non



cumulées (graphique de gauche).



- <u>Formulation du commentaire :</u> Le commentaire de ce type de graphique se fonde sur le pointage des différences et/ou ressemblances entre les deux courbes. Sur le graphique des effectifs non-cumulés, on retrouve le caractère bi-modal de la distribution observée (la distribution théorique étant nécessairement unimodale) avec des observations concentrées sur les dernières modalités (effet plafond). On observe également des écarts importants entre les deux distributions au-delà de la classe 4. On retrouve ces différences sur le graphique des effectifs cumulés puisque, sur ce graphique, les effectifs cumulés théoriques augmentent plus rapidement que les effectifs cumulés observés à partir de la 4ème classe.
  On peut donc dire que la distribution observée est éloignée d'une distribution normale, ce que confirme le calcul des écarts entre les deux distributions, avec un écart de 9,139 pour la classe 31-35.

Exercice 18.

Dans cet exercice, nous avons deux variables.

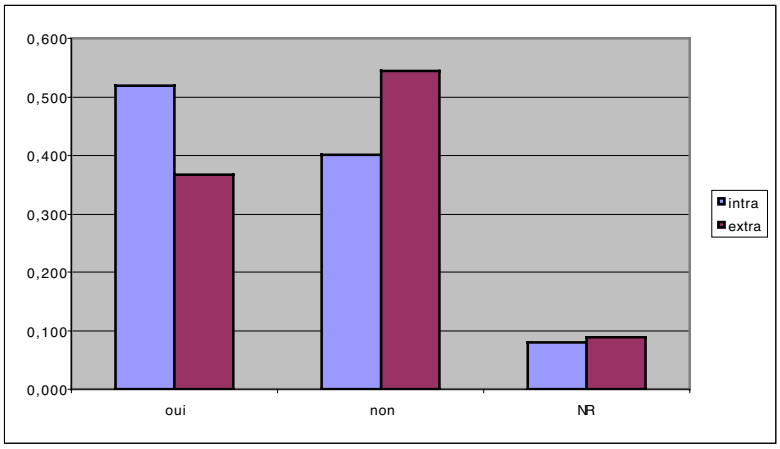
- La première est le facteur « lieu de travail ». C'est la variable indépendante. L'échelle est nominale. Elle comprend deux modalité qui vont constituer les deux groupes d'observations. Chaque sujet n'appartenant qu'à une des catégories de personnel, la structure du protocole est donc l'emboîtement (groupes indépendants).
- La seconde variable est la réponse. C'est la variable dépendante. L'échelle de mesure est une échelle nominale à trois modalités.
- Il faut donc réaliser une comparaison de deux groupes indépendants sur une variable nominale. La base de comparaison sera donc la fréquence.

Première étape : calcul des fréquences. Partant du tableau de distribution, il faut calculer les fréquences de chacune des réponses pour chacune des catégories de personnel. Attention, ce qu'on cherche à comparer ce sont les fréquences des réponses dans chacun des groupes d'observations. On doit donc calculer ces fréquences sur l'effectif total de chacun des groupes. La fréquence se calcule simplement en divisant l'effectif de la case par l'effectif total.

Ainsi pour le personnel de l'intra-hospitalier ayant répondu « oui », on aura 150/289=0,519. De la même façon, pour ceux qui ont répondu « non », on aura 116/289=0,401 etc. Le total bien sûr est égal à 1 dans chacun des groupes. On peut également exprimer ces fréquences en pourcentages.

	intra	*extra*
oui	0,519	0,367
non	0,401	0,544
NR	0,080	0,089
Total	1,000	1,000

Deuxième étape : représentation graphique. Elle n'apporte pas d'information supplémentaire, mais facilite la comparaison des groupes d'observations.



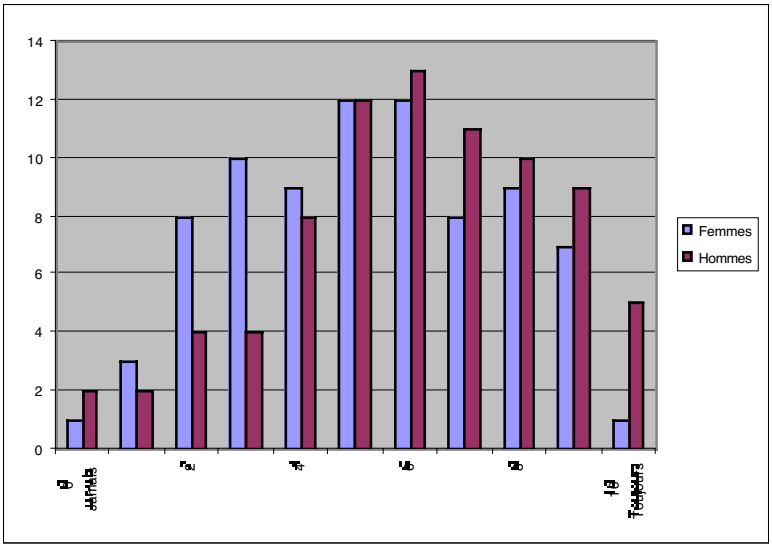
Troisième étape : commentaires. Elle consiste à pointer les principales différences entre les deux groupes et à en tirer une conclusion. On voit, dans notre exemple, que la réponse «oui » est plus fréquente pour le personnel « intra », tandis que le personnel de l'extra-hospitalier répond plus fréquemment « non ». Il n'y a pas de différence entre les groupes pour les non-réponses. On peut donc dire que le personnel de l'intra-hospitalier ressent plus une menace sur les emplois que le personnel de l'extra-hopitalier.

Exercice 19.

Dans cet exercice, nous avons deux variables. La première est la variable indépendante (facteur) « sexe ». C'est une variable nominale à deux modalités. Comme dans les exercices précédents, c'est le facteur qui sert de base à la constitution des groupes d'observations. Ces deux groupes sont, bien entendu, indépendants (relation d'emboîtement) puisque les sujets ne peuvent (sauf bizarrerie de la nature) appartenir qu'à un seul groupe.

La seconde variable est la variable dépendante « jugement ». Pour cette variable, les 11 modalités sont ordonnées, mais la notion d'intervalle n'a pas de sens, bien que les modalités soient exprimées par des chiffres. Cependant, sur ce type d'échelle de jugement de fréquences, certains auteurs franchissent le pas et considèrent qu'il existe une continuité et un intervalle entre les modalités. Ce point de vue n'est pas dénué de sens, mais demande à être justifié. Dans le doute, on peut toujours considérer une échelle numérique comme une échelle ordinale puisque la relation d'ordre est commune aux deux échelles. L'inverse n'est, bien entendu, pas vrai. Nous considèrerons donc que l'échelle de mesure est une échelle ordinale.

Notre tâche consiste ici à comparer deux groupes indépendants sur une variable ordinale. La base de la comparaison peut donc être soit les fréquences, soit la médiane et les quartiles. Dans la mesure où les effectifs des groupes sont équilibrés (même nombre de sujets dans chaque groupe), il est équivalent de comparer les effectifs ou les fréquences. On peut donc directement construire le graphique des distributions pour les deux groupes. Le graphique est le suivant :



Commentaires : On peut voir que ces deux distributions sont bi-modales (deux modes) qui correspondent dans les deux distributions aux notes 5 et 6. Les réponses se concentrent, pour les hommes et les femmes, sur un jugement moyen qui correspond à « parfois ». Cependant, on peut également noter que la distribution des réponses chez les hommes est plutôt décalée vers la droite, c'est-à-dire les notes hautes. Ce qui suggère que les hommes attribuent plus fréquemment la cause d'un accident à une défaillance mécanique.

On peut voir dans cette première analyse que ce qui différencie les deux groupes, c'est surtout la répartition des réponses sur l'échelle de mesure. Il est donc tout à fait intéressant de résumer cette répartition. Les indices pertinents sont alors la médiane et les quartiles.

Rappelons que pour les situer, il faut d'abord construire une distribution (voir les résumés d'une distribution) et ce pour chacun des deux groupes, et repérer ensuite les modalités correspondant à n/4 pour Q1, n/2 pour Q2 et n*3/4 pour Q3. Attention, ici n est le nombre total d'observations de chacun des groupes d'observations soit 80. Les distributions cumulées sont donc les suivantes :

Notes	*Femmes*	*Hommes*
0	1	2
1	4	4
2	12	8
3	22	12
4	31	20
5	43	32
6	55	45
7	63	56
8	72	66
9	79	75
10	80	80

Ce qui nous conduit à situer les quartiles de la manière suivante :

	Femmes	Hommes
Quartile 1	3	4
Médiane	5	6
Quartile 3	7	8

On peut voir sur ces résumés que les distributions sont également étendues. L'écart entre le quartile 1 et le quartile 3 est en effet de 4 pour les deux distributions. En revanche, la distribution pour les hommes est décalée vers les notes hautes. Ils considèrent donc que les accidents sont plus fréquemment dus à une défaillance mécanique que les femmes.

Exercice 20.

Dans cet exemple, il faut faire abstraction de la notion de couples pour répondre à la question. Il s'agit en fait de comparer le groupe des femmes et des hommes sur la variable « âge au moment du mariage ». Les individus statistiques sont donc les personnes. Nous avons deux variables : le sexe (variable nominale indépendante) et l'âge au moment du mariage (variable dépendante numérique). Le protocole est structuré par une relation d'emboîtement (chaque sujet est caractérisé par un seul des deux sexes). Nous avons donc à comparer deux groupes indépendants sur une variable numérique.

Une autre façon de voir ce protocole est de considérer que les individus statistiques sont les couples. Dans ce cas, chaque couple est caractérisé par l'âge de l'époux et l'âge de l'épouse au moment du mariage. Ce sont deux variables observées et il n'y a pas de facteur. Nous avons donc un protocole bivarié non structuré. Un tel point de vue sur le protocole ne permet pas de répondre à la question posée pour laquelle il faut disposer d'un protocole structuré. Il permet en revanche de répondre à une autre question : « Dans un couple, existe-il un lien entre l'âge de l'époux et l'âge de l'épouse ? ».Cette question relève de l'étude de la relation entre variables qui sera traitée au prochain chapitre.

Revenons à la question posée : « Peut-on dire que les hommes se marient plus tard que les femmes ? ». Nous pouvons utiliser comme base de comparaison la fréquence, mais l'échelle de mesure est très étendue (les observations vont de 18 à 45) ce qui va rendre difficile la comparaison à cause de la dispersion des observations sur les différentes modalités. Les quartiles sont également utilisables comme base de comparaison mais l'étendue importante de la variable rend ces indices moins discriminants. Nous choisirons donc la moyenne comme base de comparaison. Nous ne reprendrons pas ici la procédure de calcul de la moyenne, attention cependant, dans cet exercice, vous partez du protocole et non de la distribution. On peut également évaluer la dispersion des observations dans chaque groupe en calculant les écarts-types. Le résultat de ces calculs est le suivant :

	Hommes	Femmes
Moyenne	27	26
Variance	41,590	37,231
Ecart-type	6,449	6,102

Commentaires : On peut voir que les hommes se marient en moyenne un an plus tard que les femmes. La dispersion dans les deux groupes est sensiblement la même. Il n'y a pas de plus grande disparité de l'âge au moment du mariage dans l'un ou l'autre groupe.

Exercice 21.

Dans ce protocole, nous avons deux variables. La première est la catégorie professionnelle. Bien que les effectifs aient été égalisés pour faciliter les comparaisons, c'est bien une variable observée. L'échelle de mesure est nominale (considérer que les catégories sont ordonnées fait implicitement appel à d'autres variables comme que le niveau d'études ou le salaire qui ne sont pas forcément en concordance avec la catégorie professionnelle). La seconde variable est la réponse à la question I. C'est également une variable nominale observée. Nous avons donc un protocole bivarié. Les variables étant nominales, c'est le calcul de Φ^2 qui nous permettra d'analyser la liaison entre les variables.

Première étape : calcul des fréquences. Le point de départ de l'analyse est le tableau des effectifs conjoints. À partir de ce tableau, on calcule les fréquences en divisant l'effectif de la case par l'effectif total. Attention, contrairement à la comparaison de groupes, on calcule ces fréquences sur l'effectif total et non sur l'effectif du groupe.

Un peu de formalisme avant d'aller plus loin. Nous allons appeler j les modalités de la variable disposée en ligne et k les modalités de la variable disposée en colonne. La lettre f désigne toujours une fréquence et la lettre n un effectif. La fréquence d'une case est donc : $f_{jk}=n_{jk}/n$.

Concrètement : f_{11} (lire f-1-1) désigne la fréquence de la première modalité en ligne et de la première modalité en colonne c'est-à-dire les cadres supérieurs ayant répondu « parce qu'ils se sentent dans un état anormal ».

$f_{11}=n_{11}/n=1/80=0,013$; $f_{12}=n_{12}/n=7/80=0,088$ etc.

Fréquences.	*Cadre sup.*	*Cadre moy.*	*Ouvriers*	*Prof. Lib.*	*Total*
Parce qu'ils se sentent dans un état anormal.	0,013	0,088	0,038	0,075	0,213
Par besoin d'aide, de conseil.	0,125	0,075	0,025	0,075	0,300
Par besoin de se connaître.	0,075	0,038	0,100	0,038	0,250
Pour des problèmes d'orientation.	0,025	0,013	0,038	0,063	0,138
Autres réponses ou absence de réponses.	0,013	0,038	0,050	0,000	0,100
Total	0,250	0,250	0,250	0,250	1,000

On remarquera dans ce tableau que les fréquences marginales en colonne sont toutes égales. Ce qui est normal, puisque les effectifs marginaux en colonne sont égaux. Le total général des fréquences est bien sûr égal à 1.

Deuxième étape : calcul des fréquences-produits. À partir du tableau des fréquences, on va calculer, pour chaque case le produit de ses fréquences marginales (total des fréquences en ligne et en colonnes). Nous les noterons f'. On aura ainsi $f'_{jk}=f_j*f_k$.
Concrètement : f'11=0,250*0,213=0,053 ; f'12=0,250*0,300=0,053 etc.

Fréquences-produits	*Cadre sup.*	*Cadre moy.*	*Ouvriers*	*Prof. Lib.*	*Total*
Parce qu'ils se sentent dans un état anormal.	0,053	0,053	0,053	0,053	0,213
Par besoin d'aide, de conseil.	0,075	0,075	0,075	0,075	0,300
Par besoin de se connaître.	0,063	0,063	0,063	0,063	0,250
Pour des problèmes d'orientation.	0,034	0,034	0,034	0,034	0,138
Autres réponses ou absence de réponses.	0,025	0,025	0,025	0,025	0,100
Total	0,250	0,250	0,250	0,250	1,000

On remarquera que les fréquences-produits marginales sont les mêmes que dans le tableau précédent. On notera aussi que, pour une modalité de réponse particulière, les fréquences-produits sont les mêmes pour chacune des catégories professionnelles. Cela tient au fait que les effectifs marginaux en colonne sont les mêmes. En fait le calcul des fréquences-produits revient à calculer les fréquences qu'on obtiendrait si les réponses des sujets se répartissaient de la même façon pour chacune des catégories professionnelles, c'est-à-dire s'il n'y avait aucune liaison entre les variables. Dit autrement, s'il n'y a pas de relation entre les variables, la distribution des fréquences des réponses est la même pour les quatre catégories professionnelles. C'est cette distribution, que nous appellerons distribution théorique, que nous venons de calculer. Pour la comparer à la distribution observée, il nous faut connaître les effectifs correspondant à cette distribution des fréquences.

Troisième étape : calcul des effectifs théoriques. Ils s'obtiennent en multipliant les fréquences-produits par l'effectif total. L'effectif théorique se note n'. On aura donc $n'_{jk}=f'_{jk}*n$. Concrètement $n'_{11}=f'_{11}*n=0,053*80= 4$; $n'_{12}=f'_{12}*n=0,053*80$ etc.

Effectifs théoriques	*Cadre sup.*	*Cadre moy.*	*Ouvriers*	*Prof. Lib.*	*Total*
Parce qu'ils se sentent dans un état anormal	4,25	4,25	4,25	4,25	17,00
Par besoin d'aide, de conseil	6,00	6,00	6,00	6,00	24,00
Par besoin de se connaître	5,00	5,00	5,00	5,00	20,00
Pour des problèmes d'orientation	2,75	2,75	2,75	2,75	11,00
Autres réponses ou absence de réponses.	2,00	2,00	2,00	2,00	8,00
Total	20,00	20,00	20,00	20,00	80,00

On notera que les effectifs théoriques sont les mêmes pour chacune des catégories professionnelles, ce qui est normal puisque les effectifs totaux sont les mêmes pour chacune des colonnes. On notera également que les effectifs marginaux théoriques sont les mêmes que les effectifs marginaux observés. C'est toujours le cas. En effet, les effectifs théoriques correspondent à une répartition des observations proportionnelle aux effectifs marginaux, il est donc normal de retrouver les mêmes marges.

Quatrième étape : Calcul des taux de liaison. Si nos deux variables sont liées, alors les effectifs observés s'écartent de manière importante des effectifs théoriques. On évalue ces écarts en faisant simplement la différence entre les effectifs observés et les effectifs théoriques. Bien sûr ces écarts n'ont de sens que relativement aux effectifs attendus en cas d'absence de liaison. C'est la raison pour laquelle on pondère ces écarts par les effectifs théoriques. Concrètement, le taux de liaison s'obtient de la manière suivante :

$$\text{Taux de liaison} = n_{jk}-n'_{jk} /n'_{jk}.$$

Pour la première case, on aura donc (1-4,25)/4,25= -0,765 ; Pour la seconde case, on aura (7-4,25)/4,25= 0,647.

Taux de liaison	*Cadre sup.*	*Cadre moy.*	*Ouvriers*	*Prof. Lib.*
Parce qu'ils se sentent dans un état anormal.	**-0,765**	0,647	**-0,294**	0,412
Par besoin d'aide, de conseil.	0,667	0,000	**-0,667**	0,000
Par besoin de se connaître.	0,200	**-0,400**	0,600	**-0,400**
Pour des problèmes d'orientation.	**-0,273**	**-0,636**	0,091	0,818
Autres réponses ou absence de réponses.	**-0,500**	0,500	1,000	**-1,000**

Le commentaire de ce tableau s'appuiera sur le sens des écarts et l'importance de ces écarts. Les taux de liaison positifs indiquent les sur-représentations. Les taux de liaison négatifs indiquent les sous-représentations. La valeur absolue du taux de liaison révèle l'importance de l'écart entre les effectifs observés et les effectifs théoriques. On observe ici que les cadres

supérieurs ont tendance à répondre plus souvent « Par besoin d'aide, de conseil » et répondent moins souvent « Parce qu'ils se sentent dans un état anormal », contrairement aux cadres moyens. Chez les ouvriers en revanche, c'est l'absence de réponses qui prédomine, contrairement aux professions libérales qui s'abstiennent rarement de répondre et pour qui la consultation d'un psychologue relève essentiellement de problèmes d'orientation.

Cinquième étape : calcul du carré moyen de contingence Φ^2. Nous avons vu à l'étape précédente que les modalités de réponse sont liées à la catégorie professionnelle des sujets. La dernière étape consiste à évaluer globalement l'importance de cette liaison. Cette évaluation se fait à l'aide de la statistique Φ^2. Pour chaque case, on calcule le carré du taux de liaison qu'on multiplie ensuite par la fréquence-produit correspondante. Concrètement pour la première case (cadres sup./réponse 1), on aura : $-0{,}765^2*0{,}053= 0{,}031$; Pour la seconde case, on aura : $0{,}647^2*0{,}053= 0{,}022$.

Carré moyen de contingence	*Cadre sup.*	*Cadre moy.*	*Ouvriers*	*Prof. Lib.*	*Total*
Parce qu'ils se sentent dans un état anormal.	0,031	0,022	0,005	0,009	0,067
Par besoin d'aide, de conseil.	0,033	0,000	0,033	0,000	0,067
Par besoin de se connaître.	0,003	0,010	0,023	0,010	0,045
Pour des problèmes d'orientation.	0,003	0,014	0,000	0,023	0,040
Autres réponses ou absence de réponses.	0,006	0,006	0,025	0,025	0,063
Total	0,076	0,052	0,086	0,067	0,281

Φ^2 est égal au total du tableau. Dans cet exemple, Φ^2=0,281.

La méthode que nous venons de présenter est celle de votre cours, en pratique on ne procède pas tout à fait de cette façon. On ne construit qu'un seul tableau, en organisant les cellules de la façon suivante :

Effectif observé	Effectif théorique
Ecart brut	Contribution au X^2

On peut ainsi gagner tout le temps nécessaire à la recopie des tableaux, et avoir sous les yeux de manière synthétique tous les résultats de l'analyse. L'autre intérêt, c'est que, pour chacune des cellules du tableau, on enchaîne les calculs (il est donc inutile de saisir à nouveau le résultat intermédiaire) et si on sait se servir de la mémoire de sa calculette, on peut calculer parallèlement le X^2 qu'il ne reste plus qu'à diviser par l'effectif total pour avoir Φ^2.

Concrètement, voici comment on s'y prend :

Dans les marges du tableau, on note les effectifs totaux observés et dans le coin en haut à gauche de chaque cellule, on note les effectifs observés. On calcule ensuite les contributions au X2 de chaque cellule de la manière suivante :

Exemple : Pour les cadres supérieurs ayant donné la première modalité de réponse :

- ✓ On calcule l'effectif théorique en multipliant les marges et en la divisant par n soit 17*20/80=4,25. On note ce résultat en haut à droite de la cellule.
- ✓ On calcule ensuite l'écart brut. Pour ne pas avoir à saisir à nouveau l'effectif théorique sur la calculette, on soustrait à l'effectif théorique l'effectif observé et on inverse le signe. Dans notre exemple : 4,25-1=3,25. l'écart brut est donc de –3,25. On le note en bas à gauche dans la cellule.
- ✓ Puis on élève au carré l'écart brut . Le résultat est ensuite divisé par l'effectif théorique. On note le résultat de cette dernière opération dans le coin en bas à droite.
- ✓ Si votre calculette dispose d'une touche mémoire (notée généralement « M+ »), on additionne ce dernier résultat au contenu de la mémoire.
- ✓ Puis on recommence pour chacune des cellules du tableau. Lorsque vous aurez fini, la mémoire de votre calculette devrait contenir la somme des contributions au X^2, c'est-à-dire X^2. On divise alors ce résultat par n pour obtenir Φ^2. Dans notre exemple, X^2=22,468 et donc Φ^2=22,468/80=0,281

	Profession								
Réponses	***Cadre sup.***		***Cadre moy.***		***Ouvriers***		***Prof. Lib.***		***total***
Parce qu'ils se sentent dans un état anormal.	1	4,25	7	4,25	3	4,25	6	4,25	17
	-3,25	2,485	2,75	1,779	-1,25	0,368	1,75	0,721	
Par besoin d'aide, de conseil.	10	6	6	6	2	6	6	6	24
	4	2,667	0	0	-4	2,667	0	0	
Par besoin de se connaître.	6	5	3	5	8	5	3	5	20
	1	0,2	-2	0,8	3	1,8	-2	0,8	
Pour des problèmes d'orientation.	2	2,75	1	2,75	3	2,75	5	2,75	11
	-0,75	0,205	-1,75	1,114	0,25	0,023	2,25	1,841	
Autres réponses ou absence de réponses.	1	2	3	2	4	2	0	2	8
	-1	0,5	1	0,5	2	2	-2	2	
Total	20		20		20		20		80

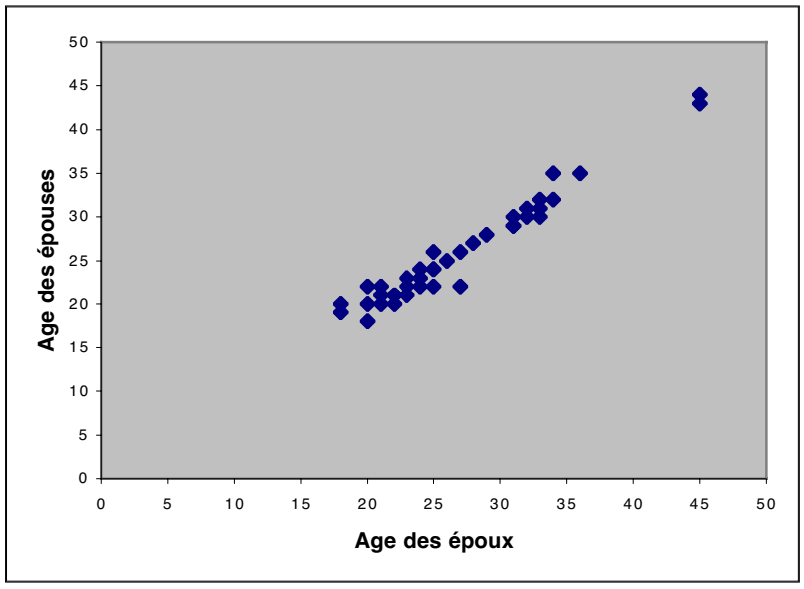
Interprétation du résultat.

Cet indice Φ^2 peut être vu comme la proportion de sujets s'écartant du cas d'absence de liaison. Sa valeur varie entre 0 (absence de liaison) et 1 (concordance entre les modalités des

variables). Dans notre exemple, Φ^2 est plus proche de 0 que de 1, nous conclurons donc à une liaison faible entre les variables. Le cas de concordance correspond au cas où, pour chaque catégorie professionnelle, on aurait observé qu'une seule modalité de réponse. Dans un tel cas, on comprend bien que connaissant la réponse d'un sujet à la question, on connaît également sa catégorie professionnelle. De la même façon, si je connais la catégorie professionnelle d'un sujet, je sais, dans le cas de concordance, ce qu'il va répondre à la question. Ce cas n'est bien sûr que très rarement observé, mais s'il existe une liaison entre les variables, c'est de cela qu'on se rapprocherait. On pourra, à titre d'exercice, calculer Φ^2 pour le cas de concordance et vérifier que sa valeur est de 1. De la même façon, si la liaison entre les variables est nulle, le tableau des effectifs observés serait le même que celui des effectifs théoriques. On peut alors vérifier que Φ^2 est bien égal à 0.

Exercice 22.

La question que l'on se pose ici est très différente de celle que nous avons examinée dans le chapitre précédent. La question de la relation entre les variables demande de voir le protocole différemment. Ici, les unités statistiques sont les couples et nous avons deux variables numériques : l'âge de l'époux et l'âge de l'épouse. On considère donc ici un protocole bivarié non structuré. L'évaluation de la liaison se fera à l'aide du r de Bravais-Pearson. Celui-ci est en effet un bon indice pour évaluer la liaison entre les deux variables puisqu'on peut voir sur le graphique ci-dessous que les données suivent à peu près une droite.



L'application de la formule de définition nécessite le calcul des quantités :

- $\sum (x_i - m_x)(y_i - m_y)$
- $\sum (x_i - m_x)^2$
- $\sum (y_i - m_y)^2$

La première s'obtient en calculant pour chaque couple et chaque variable l'écart entre l'observation et la moyenne et en faisant ensuite le produit de ces écarts.

Par exemple, pour le premier couple, on aura : (20-27)*(20-26)=-7*-6=42 ;

Pour le second couple, on aura (25-27)*(24-26)=-2*-2=4.

Lorsque cela est réalisé pour chacun des couples, on fait la somme de ces produits. Ici le total est égal à 1501.

Pour les deux autres quantités, on reconnaît la formule du dénominateur de la variance. La procédure de calcul est la même. En pratique, les écarts à la moyenne ayant été calculés pour la précédente quantité, on les élève au carré et on fait ensuite la somme de ces carrés pour chacune des variables. Dans notre exemple, nous avons :

$$\sum (x_i - m_x)^2 = 1622 \text{ et } \sum (y_i - m_y)^2 = 1452$$

On applique ensuite la formule de définition pour obtenir le r de Bravais-Pearson :

$$r = \frac{\sum (x_i - m_x)(y_i - m_y)}{\sqrt{\sum (x_i - m_x)^2 \sum (y_i - m_y)^2}} = \frac{1501}{\sqrt{1622 * 1452}} = 0,978$$

Couples	*Epoux*	*Epouse*	x_i-m_x	y_i-m_y	$(x_i$-$m_x)$*$(y_i$-$m_y)$	$(x_i$-$m_x)^2$	$(y_i$-$m_y)^2$
1	20	20	-7	-6	42	49	36
2	25	24	-2	-2	4	4	4
3	25	22	-2	-4	8	4	16
4	31	29	4	3	12	16	9
5	18	20	-9	-6	54	81	36
6	25	24	-2	-2	4	4	4
7	32	31	5	5	25	25	25
8	26	25	-1	-1	1	1	1
9	22	21	-5	-5	25	25	25
10	24	24	-3	-2	6	9	4
11	24	22	-3	-4	12	9	16
12	25	26	-2	0	0	4	0
13	27	22	0	-4	0	0	16
14	20	18	-7	-8	56	49	64
15	18	19	-9	-7	63	81	49
16	23	21	-4	-5	20	16	25
17	23	23	-4	-3	12	16	9
18	34	35	7	9	63	49	81
19	45	43	18	17	306	324	289
20	23	22	-4	-4	16	16	16
21	22	21	-5	-5	25	25	25
22	20	22	-7	-4	28	49	16
23	21	22	-6	-4	24	36	16
24	34	32	7	6	42	49	36
25	21	20	-6	-6	36	36	36
26	32	30	5	4	20	25	16
27	45	44	18	18	324	324	324
28	33	32	6	6	36	36	36
29	31	30	4	4	16	16	16
30	33	31	6	5	30	36	25
31	36	35	9	9	81	81	81
32	28	27	1	1	1	1	1
33	27	26	0	0	0	0	0
34	29	28	2	2	4	4	4
35	21	21	-6	-5	30	36	25
36	22	20	-5	-6	30	25	36
37	24	23	-3	-3	9	9	9
38	31	29	4	3	12	16	9
39	33	30	6	4	24	36	16
				Total	1501	1622	1452

Interprétation.

La valeur de r est très proche de 1. On peut donc affirmer qu'il existe une relation linéaire positive entre l'âge de l'époux et l'âge de l'épouse dans un couple. Dit autrement, les âges des conjoints dans un couple sont proportionnels.

Pour l'application de la formule de calcul, on se reportera au cours qui la présente de façon détaillée. On pourra à titre d'exercice la mettre en œuvre et vérifier qu'on obtient bien le même résultat (aux arrondis près).

Une autre façon de calculer le r de Bravais-Pearson est de calculer la covariance et de la diviser par le produit des deux écarts-types. La covariance est la moyenne des produits des écarts à la moyenne. Formellement, la covariance se définit ainsi[6] :

$$\mathrm{cov}_{xy} = \frac{\sum (x_i - m_x)(y_i - m_y)}{n}$$

La formule du r de Bravais-Pearson peut donc se réécrire de la façon suivante :

$$r = \frac{\mathrm{cov}_{xy}}{s_x * s_y} = \frac{\frac{\sum (x_i - m_x)(y_i - m_y)}{n}}{\sqrt{\frac{\sum (x_i - m_x)^2 \sum (y_i - m_y)^2}{n}}}$$

On identifie aisément la simplification qui permet de passer à la formule de définition du r de Bravais-Pearson. Cette procédure de calcul est surtout intéressante si on connaît déjà les écarts-types.

Concrètement, il faut :

- ✓ Calculer pour chaque individu statistique et chaque variable, son écart à la moyenne. La moyenne pour l'âge des époux étant de 27 ans et la moyenne de l'âge des épouses étant de 26 ans, on a pour le premier couple un écart à la moyenne de 20-27=-7 pour l'âge de l'époux et un écart de 20-26=-6 pour l'âge de l'épouse.
- ✓ On fait ensuite le produit de ces deux écarts. Ainsi, pour le premier couple, on aura : -7*-6=42.
- ✓ On procède ainsi pour tous les couples (voir le tableau ci-dessous).
- ✓ On calcule ensuite la moyenne des produits des écarts aux moyennes (concrètement la moyenne de la dernière colonne). Cette moyenne est la covariance. Dans notre exemple, elle vaut 38,487.
- ✓ On divise ensuite cette covariance par le produit des deux écarts-types. On a donc :

$$r = \frac{\mathrm{cov}_{xy}}{s_x * s_y} = \frac{38,487}{6,149 * 6,102} = 0,978$$

[6] Attention, n représente le nombre d'individus statistiques et non le nombre d'observations. Nous avons 39 couples et 39*2=78 observations puisque nous avons deux variables.

Couples	*Epoux*	*Epouse*	*xi-mx*	*yi-my*	*(xi-mx)*(yi-my)*
1	20	20	-7	-6	42
2	25	24	-2	-2	4
3	25	22	-2	-4	8
4	31	29	4	3	12
5	18	20	-9	-6	54
6	25	24	-2	-2	4
7	32	31	5	5	25
8	26	25	-1	-1	1
9	22	21	-5	-5	25
10	24	24	-3	-2	6
11	24	22	-3	-4	12
12	25	26	-2	0	0
13	27	22	0	-4	0
14	20	18	-7	-8	56
15	18	19	-9	-7	63
16	23	21	-4	-5	20
17	23	23	-4	-3	12
18	34	35	7	9	63
19	45	43	18	17	306
20	23	22	-4	-4	16
21	22	21	-5	-5	25
22	20	22	-7	-4	28
23	21	22	-6	-4	24
24	34	32	7	6	42
25	21	20	-6	-6	36
26	32	30	5	4	20
27	45	44	18	18	324
28	33	32	6	6	36
29	31	30	4	4	16
30	33	31	6	5	30
31	36	35	9	9	81
32	28	27	1	1	1
33	27	26	0	0	0
34	29	28	2	2	4
35	21	21	-6	-5	30
36	22	20	-5	-6	30
37	24	23	-3	-3	9
38	31	29	4	3	12
39	33	30	6	4	24
Total	1053	1014		Covariance	38,487

Bibliographie.

Plus de cours ?

Beaufils B. (1998) Statistiques appliquées à la psychologie : tome 1, Statistiques descriptives. Bréal, collection " lexifac ".

Howell D.C. (1998) Méthodes statistiques en sciences humaines, Bruxelles : De Boeck Université, CollectionMéthodes en sciences humaines.

Rouanet H., Le Roux B., Bert M.C. (1987) Statistique en sciences humaines, procédures naturelles, Paris, Dunod.

Plus d'entrainement ?

Beauvois J.L., Dubois N. (1998) Exercices de psychologie, tome 2, bases et méthodes-, Paris, Dunod, collection " Psycho sup ".

Gueguen N. (2001) Statistique pour psychologue : Cours et exercices, Paris, Dunod, collection " Psycho sup ".

Rouanet H., Le Roux B. (1995) Statistiques en sciences humaines : Exercices et solutions, Paris, Dunod, collection " Psycho sup ".

Et des liens...

Encore des cours (en français) !

Intitiation aux méthodes statstistiques :
http://ibm2.cicrp.jussieu.fr/grasland/STAT98/STAT98.htm

SEL : http://www.inrialpes.fr/sel/telecharger.html

Statnet : http://www.agro-montpellier.fr/cnam-lr/statnet/

Mas11 : http://nte-serveur.univ-lyon1.fr/nte/immediato/math2002/Mass11/cours/cours_table.htm

Fiches de statistique : http://hpa.free.fr/Fiches_de_Stat.htm

Encore des cours (en anglais)!

HyperStat Online : http://davidmlane.com/hyperstat/index.html

The Statistics Homepage : http://www.statsoft.com/textbook/stathome.html

StatNotes: http://www2.chass.ncsu.edu/garson/pa765/statnote.htm

Introductory Statistics: http://www.psychstat.smsu.edu/introbook/sbk00.htm

Et si l'ordinateur vous aidait ?

StatView 5 Demos & Downloads : http://www.statview.com/product/demo.shtml

Formulaire

Statistiques	Formules
Densité	$Densité = effectif / étendue$
Fréquence	$f(u) = \frac{n(u)}{n}$
Moyenne	$m = \frac{\sum x_i}{n} = \frac{T}{n} = \frac{\sum n_k u^k}{n} = \frac{\sum x_i n_i}{n}$
Variance	$Var = \frac{\sum (x_i - m)^2}{n} = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n}$
Ecart-type	$Ecart - type = \sqrt{\text{var}}$
Variance corrigée	$Var_{corr} = \frac{\sum (x_i - m)^2}{n-1} = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} = Var * n / n - 1$
Ecart-type corrigé	$Ecart - type_{corr} = \sqrt{\text{var}_{corr}}$
Note z	$z_i = \frac{x_i - m}{écart - type}$
Effectifs théoriques	$n\mathbb{C}_{jk} = \frac{n_j * n_k}{n}$
Taux de liaison	$Taux.de.liaison = \frac{n_{jk} - n\mathbb{C}_{jk}}{n\mathbb{C}_{jk}}$
Khi-deux	$X^2 = \sum \frac{(obs - théo)^2}{théo} = \sum \frac{(n_{jk} - n\mathbb{C}_{jk})^2}{n\mathbb{C}_{jk}} = \sum \frac{\left(n_{jk} - \frac{(n_j * n_k)}{n}\right)^2}{\frac{(n_j * n_k)^2}{n}}$
Carré moyen de contingence	$\Phi^2 = \sum \frac{(n_{jk} - n\mathbb{C}_{jk})^2}{(n\mathbb{C}_{jk})^2} * \frac{n\mathbb{C}_{jk}}{n} = \frac{\chi^2}{n}$
Coefficient de corrélation r de Bravais-Pearson	$r = \frac{\sum (x_i - m_x)(y_i - m_y)}{\sqrt{\sum (x_i - m_x)^2 \sum (y_i - m_y)^2}} = \frac{\sum x_i y_i - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x_i^2 - \left(\sum x_i\right)^2 / n\right)\left(\sum y_i^2 - \left(\sum y_i\right)^2 / n\right)}}$
Coefficient de corrélation par rang de Spearman	$\rho = 1 - \frac{6 \sum d^2}{n^3 - n}$