STAT 4 STAPS

Christophe Genolini

28 avril 2006

Table des matières

Chapitre 1

Introduction

1.1 Les stats, pour quoi faire ?

Dans son besoin de comprendre le monde, l'homme cherche en permanence à établir des règles : un objet plus léger que l'eau flotte (Archimède) ; la terre tourne autour du soleil (Galilée) ; une pomme qui se détache de l'arbre tombe vers le sol (Newton) ; rien n'est plus rapide que la lumière (Einstein)...

Malheureusement, toutes les lois de la nature ne sont pas toutes aussi strictes. Un verre qui tombe se casse... souvent, mais pas toujours. Au tennis, le 400^{e} joueur mondial affronte le premier. Il va perdre. A moins que ?

En Sciences de la vie, les règles absolues sont encore plus difficiles à établir. D'un individu à l'autre, les mêmes causes ne produisent pas les mêmes effets. Un dépressif sera soigné par un médicament qui restera sans effet sur un autre. Certain fumeurs attrapent le cancer, d'autres non. Là ou ça se complique, c'est que des non fumeurs attrapent aussi le cancer. Que peut-on en conclure ? Il est clair qu'aucune règle absolue du genre "fumer = cancer, ne pas fumer = bonne santé" ne peut être établie. Faut-il pour autant renoncer à établir un lien entre la cigarette et le cancer ? Non. Puisqu'une règle absolue n'existe pas, on peut tout de même essayer de trouver une règle probable, quelque chose du genre : "un fumeur a x% de chance de contracter un cancer des poumons, un non fumeur a y% de chances". Ca ne veut pas dire que tous les fumeurs vont tomber malades, ni qu'ils vont tous être en bonne santé. Juste qu'ils ont plus de chances que les autres. Reste à trouver le moyen d'établir ce genre de règle.

Et c'est là que les statistiques commencent...

1.2 Stat, stat et stat

Il y a trois types d'analyse statistique :

- **Résumé des données** : c'est la plus commune, celle qu'on trouve dans nos journaux. Elle décrit simplement les données. Le calcul de moyenne, d'écart type, de pourcentages et leur représentation graphique entrent dans cette catégorie.
- **Test statistique** : un test permettra de répondre à une question : "Y a-t-il plus de cancer chez ceux qui fument ?" Les tests statistiques tranchent... avec une fiabilité variable, mais que l'on peut connaître.
- **Modélisation** : enfin, les statistiques permettent de décrire le monde de manière raisonnablement précise et de jouer aux devinettes. On constate que, contrairement à une idée reçue, les gens ne sont pas si uniques que ça. Sur les trois dernières années, aucun des étudiants qui ont entre 10 et 11 en licence n'ont eu le CAPEPS. On peut donc "parier" raisonnablement que cette année, il en sera de même et donc déconseiller cette voie à ceux qui n'ont pas eu de mention du tout. Naturellement, rien n'est sûr et l'on peut se tromper, mais statistiquement, c'est rare. De même, on peut modéliser les fluctuations de la bourse en fonction des indices, les embouteillages selon les accidents, la guérison d'un cancer des testicules selon l'âge et la date de détection...

En pratique, toutes les études commencent toujours par un résumé des données. Il permet entre autre de vérifier quelques propriétés basiques de nos données, éventuellement de détecter des valeurs aberrantes et de déterminer les types d'analyses que l'on aura le droit de faire ultérieurement.

1.2.1 Résumé les données

En 1972, un chercheur du nom de Harris [Harris72] se livre à une petite expérience toute simple : il demande de l'argent à des passants. Avant de lire la suite, posez-vous la question : si vous êtes dans la rue et que vous avez un besoin urgent d'un euro, comment l'obtenir ? Il est clair qu'aucune règle du style "En demandant poliment, on obtient un euro" existe. Et pourtant, Harris trouve presque ce genre de règle... A un certain nombre de passant, il demande un dollar. A d'autres, il demande l'heure, puis quand on lui a répondu, il demande un dollar. Au total, il a contacte 200 passants pour le résultat suivant[1] :

[Id]	[Question]	[Resultat]
P1	$ seul	Non
P2	$ seul	Non
P3	Heure + $	Non
P4	$ seul	Non
P5	$ seul	Non
P6	Heure + $	Oui
P7	Heure + $	Non
P8	$ seul	Non
P9	$ seul	Non
P10	Heure + $	Oui
P11	$ seul	Non
P12	$ seul	Non
P13	Heure + $	Oui
P14	Heure + $	Non
P15	Heure + $	Non
P16	Heure + $	Non
P17	Heure + $	Non
P18	Heure + $	Oui
P19	Heure + $	Non
P20	Heure + $	Oui
P21	$ seul	Oui
P22	$ seul	Non
P23	Heure + $	Oui
P24	Heure + $	Oui
P25	$ seul	Non
P26	Heure + $	Oui
P27	$ seul	Non
P28	Heure + $	Oui
P29	$ seul	Non
P30	$ seul	Non
P31	$ seul	Non
P32	$ seul	Non
P33	$ seul	Non
P34	$ seul	Non
P35	Heure + $	Non
P36	$ seul	Oui
P37	Heure + $	Oui
P38	Heure + $	Non
P39	Heure + $	Oui
P40	Heure + $	Oui
P41	$ seul	Non
P42	Heure + $	Oui
P43	Heure + $	Non
P44	$ seul	Non
P45	$ seul	Non
P46	Heure + $	Non
P47	$ seul	Non
P48	$ seul	Non
P49	$ seul	Non
P50	$ seul	Non

[Id]	[Question]	[Resultat]
P51	Heure + $	Oui
P52	$ seul	Non
P53	$ seul	Non
P54	Heure + $	Non
P55	Heure + $	Non
P56	Heure + $	Non
P57	Heure + $	Non
P58	$ seul	Non
P59	Heure + $	Non
P60	$ seul	Non
P61	Heure + $	Non
P62	Heure + $	Non
P63	Heure + $	Non
P64	Heure + $	Oui
P65	Heure + $	Oui
P66	$ seul	Non
P67	Heure + $	Non
P68	Heure + $	Non
P69	$ seul	Non
P70	Heure + $	Oui
P71	$ seul	Non
P72	Heure + $	Oui
P73	$ seul	Non
P74	$ seul	Non
P75	$ seul	Non
P76	$ seul	Non
P77	Heure + $	Non
P78	$ seul	Non
P79	Heure + $	Oui
P80	Heure + $	Oui
P81	$ seul	Oui
P82	Heure + $	Non
P83	$ seul	Non
P84	$ seul	Non
P85	$ seul	Non
P86	Heure + $	Oui
P87	$ seul	Non
P88	$ seul	Non
P89	Heure + $	Non
P90	$ seul	Non
P91	Heure + $	Non
P92	Heure + $	Non
P93	$ seul	Non
P94	Heure + $	Oui
P95	$ seul	Non
P96	$ seul	Non
P97	$ seul	Oui
P98	Heure + $	Non
P99	$ seul	Non
P100	$ seul	Non

[Id]	[Question]	[Resultat]
P101	$ seul	Non
P102	Heure + $	Oui
P103	$ seul	Oui
P104	Heure + $	Oui
P105	Heure + $	Non
P106	$ seul	Non
P107	$ seul	Non
P108	Heure + $	Oui
P109	Heure + $	Non
P110	Heure + $	Oui
P111	Heure + $	Non
P112	Heure + $	Oui
P113	Heure + $	Non
P114	$ seul	Non
P115	Heure + $	Oui
P116	Heure + $	Non
P117	Heure + $	Non
P118	$ seul	Non
P119	Heure + $	Non
P120	Heure + $	Oui
P121	Heure + $	Non
P122	Heure + $	Oui
P123	Heure + $	Non
P124	$ seul	Non
P125	$ seul	Non
P126	Heure + $	Oui
P127	$ seul	Non
P128	$ seul	Non
P129	$ seul	Non
P130	$ seul	Non
P131	Heure + $	Non
P132	$ seul	Non
P133	Heure + $	Oui
P134	Heure + $	Non
P135	Heure + $	Non
P136	$ seul	Oui
P137	$ seul	Non
P138	Heure + $	Non
P139	Heure + $	Oui
P140	Heure + $	Oui
P141	Heure + $	Oui
P142	$ seul	Non
P143	$ seul	Non
P144	Heure + $	Non
P145	$ seul	Non
P146	$ seul	Non
P147	$ seul	Non
P148	Heure + $	Non
P149	Heure + $	Non
P150	Heure + $	Non

[Id]	[Question]	[Resultat]
P151	Heure + $	Oui
P152	$ seul	Non
P153	Heure + $	Oui
P154	$ seul	Non
P155	$ seul	Non
P156	Heure + $	Oui
P157	$ seul	Non
P158	$ seul	Non
P159	$ seul	Non
P160	Heure + $	Oui
P161	Heure + $	Non
P162	$ seul	Non
P163	Heure + $	Non
P164	$ seul	Non
P165	Heure + $	Non
P166	$ seul	Oui
P167	Heure + $	Non
P168	$ seul	Non
P169	$ seul	Non
P170	Heure + $	Oui
P171	$ seul	Non
P172	$ seul	Non
P173	$ seul	Oui
P174	$ seul	Non
P175	$ seul	Non
P176	Heure + $	Non
P177	$ seul	Non
P178	$ seul	Non
P179	$ seul	Non
P180	Heure + $	Non
P181	Heure + $	Non
P182	Heure + $	Oui
P183	Heure + $	Oui
P184	$ seul	Non
P185	$ seul	Non
P186	$ seul	Oui
P187	Heure + $	Non
P188	$ seul	Non
P189	$ seul	Non
P190	Heure + $	Non
P191	$ seul	Non
P192	$ seul	Non
P193	$ seul	Non
P194	Heure + $	Non
P195	Heure + $	Non
P196	Heure + $	Non
P197	Heure + $	Non
P198	Heure + $	Oui
P199	$ seul	Non
P200	$ seul	Non

TAB. 1.2.1.a – Harris 1972, données brutes

Que peut-on conclure de ces données ? Rien... Du moins, rien sous cette forme.

Pour essayer de dégager une tendance, regroupons les données. Mettons ensemble les personnes à qui Harris a demandé l'heure et qui ont donné un dollar, celles à qui il a demandé l'heure et qui lui ont refusé le dollar, celles à qui il a demandé simplement un dollar et qui ont accepté, celles à qui il a demandé un dollar et qui ont refusé :

Groupe	**Nombre de personne**
Heure + $, acceptent	43
Heure + $, refusent	57
$ seul, acceptent	9
$ seul, refusent	91

TAB. 1.2.1.b – Harris 1972, données triées

Parmi les 100 personnes à qui Harris a demandé l'heure puis un $, 43 ont accepté : d'où 43% de réussite. Parmi les 100 personnes à qui Harris a simplement demandé un $, 9 ont accepté : d'où 9% de réussite.

[1] Nous ne présentons pas ici l'expérience et les données originales mais des données fictives aboutissant toutefois au même résultat.

Les données s'éclairent d'un nouveau jour. On pourrait résumer ainsi : "Si, avant de demander un dollar, on demande préalablement l'heure, on augmente considérablement nos chances d'obtenir satisfaction". Surprenant, n'est ce pas ?

Cet exemple illustre bien l'intérêt d'un résumé : à partir des données recueillies (tablea1.2.1.a), on ne pouvait pas dire grand-chose. En réorganisant nos données pour leur donner plus de clarté (table 1.2.1.b), on obtient deux chiffres qui nous donnent une vision globale de l'ensemble. Ici, la conclusion est "En demandant l'heure puis un dollar, on a plus de chance de l'obtenir qu'en demandant seulement un euro", conclusion qui en soit n'a que peu d'intérêt mais qui permet par la suite aux psychologues de bâtir des théories sur le fonctionnement de l'être humain.

Au passage, cet exemple illustre aussi l'intérêt des statistiques en général : si aucune règle du style "En demandant de l'argent de telle manière, vous êtes sur d'en obtenir" ne peut être établie, on a néanmoins montré que "Demander de telle manière augmente vos chances"

1.2.2 Test statistique

L'expérience de Harris est facilement reproductible. Vous pouvez vous-même la faire : un crayon, un papier pour noter les résultats, une bonne dose de courage, puis vous descendez dans la rue et vous posez vos questions. Supposons que vous demandiez l'heure plus un euro à 100 personnes. Combien vont accepter ? Harris a trouvé 43. Mais il est fort peu probable que vous trouviez vous aussi 43. Peut-être 40. Ou 44. Ou 41...
A quoi sont dues ces différences ? Au temps ? A l'humeur de l'expérimentateur ce jour là ? Aux journaux du matin ? A l'époque de l'année ? On ne sait pas. On se contente de les attribuer au hasard. Ce hasard nous interdit d'être précis (d'ailleurs, dans son article, Harris ne dit pas 43% mais environ 43%. Le hasard a fait qu'il a trouvé 43, mais ce même hasard aurait très bien pu lui donner 41 ou 44).
Lançons-nous dans une nouvelle expérience, l'impact de la politesse : toujours dans la rue, un expérimentateur demande un euro à cent personnes. A cent autres, il demande un euro *s'il vous plait*. Dans le premier groupe, il obtient 11 de succès. Dans le deuxième, il obtient 13. Peut-il affirmer que la deuxième méthode est plus efficace que la première ? Autrement dit, la différence observée entre 11 et 13 est-elle une vraie différence liée à la formulation de la demande, ou une fausse différence due au hasard ?
C'est à cette question que permet de répondre un test statistique. Plus précisément, un test permet de quantifier le hasard, de dire : "Il y a x% de chances que la différence que vous observez soit liée au hasard". Dans notre exemple, un test nous permettra bientôt d'affirmer : "Il y a 66,34% de chances que la différence entre 11 et 13 soit liée au hasard". Le même test permettra à Harris de dire "Il y a 0,0000002% de chances que la différence entre 43 et 9 soit liée au hasard". Sous réserve que les études aient été bien menées[2], nous pourrions donc affirmer : "Etre poli ne sert pas à grand-chose ; pour obtenir des sous, il vaut surtout demander l'heure avant."

1.2.3 Modélisation

Modéliser, c'est proposer une description du monde qui soit plus simple que la réalité. La modélisation à plusieurs buts, l'un d'entre eux est de deviner ce que l'on ne connaît pas. Pour cela, on regarde ce que l'on connaît et on le généralise. Dans [Chuine04], Isabelle Chuine nous propose un exemple ou une modélisation s'avère particulièrement efficace. Elle travaille (entre autres choses) sur l'évolution du climat, en particulier sur la canicule de l'été 2003. Elle se demande si une telle chaleur s'est déjà produite ou si le phénomène présente un caractère exceptionnel. Pour cela, elle récupère les températures des étés passés. Malheureusement, les registres de la météo ne remontent pas bien loin dans le temps, elle ne trouve que les années 1750 à 2003. D'où son recours à une modélisation : sur toutes les années dont elle dispose, elle remarque que la température de l'été est très fortement liée (presque à la journée près) à la date des vendanges ! Plus il a fait chaud, plus les vendanges sont précoces[3]. Naturellement, cela marche aussi dans l'autre sens : autrement dit, à partir de la date des vendanges, elle peut "deviner" de manière très précise la température de l'été. Or, les viticulteurs Bourguignons étant des gens particulièrement soigneux, les archives des récoltes remontent à 1370...[4]

[2] C'est-à-dire exempte de biais. Un biais est un élément qui fausse l'expérience. Par exemple, si l'expérimentateur ajoute "s'il vous plait" d'un air totalement blasé ou agressif, cela fausse l'expérience.

[3] Cela semble logique, mais encore fallait-il y penser

[4] Madame Chuine a eu la bonté d'apporter une délicieuse précision sur le *soin naturel* des Bourgignons. Je cite : "Les dates de vendanges étaient tenues dans les archives paroissiales puis municipales puis préfectorales, elles faisaient l'objet de

La modélisation statistique a donc permis à Chuine de "deviner" les températures des étés entre 1370 et 1750 pour finalement conclure : "L'été 2003 présente un caractère sans précédent. La température était supérieure de +5,86°C aux normales saisonnières[5], température jamais atteinte entre 1370 et 2003. Le précédent record datait de 1523, un simple +4,10°C..."

1.3 Échantillon et population

Le manière orthogonale avec la classification que nous venons de présenter, on distingue deux types d'études :

- En **Statistiques descriptives**, le statisticien dispose de toutes les données dont il a besoin et il les analyse (selon les modes présentés plus haut).
- A l'opposé, en **Statistiques inférentielles**, la population sur laquelle on aimerait travailler est trop vaste, le recueil des données n'est pas possible. Le statisticien est donc réduit à collecter les données sur seulement une petite partie de la population. Il fait ensuite ses analyses puis *généralise* ses données à l'ensemble de la population. C'est le principe du sondage.

En pratique, Harris a croisé 200 personnes. Certes il a augmenté sa connaissance de ces personnes et il a une idée raisonnable de leur manière de réagir. Mais cela est-il vraiment intéressant ? Relativement peu... Harris n'a pas pris les noms des gens, il ne les croisera probablement plus jamais. La seule chose qui compte pour lui, c'est que ces gens étaient monsieur et madame tout le monde. Harris peut donc **généraliser** ses résultats à une population plus large. Pour cela, il a pris soin, lorsqu'il a constitué son échantillon, de choisir 200 Américains typiques et représentatif de l'Amérique, de toutes classes et de toutes conditions. Il peut donc généraliser ses résultats à tous les Américains. S'il avait fait son enquête dans une province profonde en milieu agricole en se limitant aux personnes âgées, il n'aurait pu généraliser qu'aux agriculteurs retraités, parce qu'il est probable que les agriculteurs et les citadins ont un comportement différent, tout comme les retraités et les actifs.

Le résultat de Harris n'a donc de valeur que s'il est généralisable. Beaucoup de résultats sont dans ce cas, en particulier toutes les recherches médicales. Savoir que dans l'hôpital [[[première mise en évidence du lien cancer-tabac]]], 80% des cancer des poumons touchaient les fumeurs n'avait en soit aucun intérêt. Mais le fait que ce résultat soit généralisable à la population a un impact considérable en terme de santé publique : c'est là qu'on a découvert que fumer tue ; pas seulement les patients de l'hôpital mais tout le monde...

Chuine travaille dans un contexte complètement différent. Elle a récolté des données, les a décrites et a fabriqué un modèle. Suite à cela, elle dispose de toutes les données dont elle a besoin. Elle ne travaille pas sur un sous-ensemble d'années mais sur toutes les années possibles relativement à son modèle. Elle n'a donc pas besoin d'étendre ses résultats à un ensemble plus vaste, ils sont en eux-mêmes intéressants : la température que nous avons subit l'an dernier était exceptionnelle...

Les deux classifications sont orthogonales, on peut trouver des exemples dans chacun des cas :

---

bans (bans des vendanges) qui étaient publiés. Cela devait être respecté **sous peine d'amende**!" Grande déception, les Bourguignons n'étaient pas spécialement plus soigneux que les autres, ils étaient juste menacés financièrement...

[5] Pour information, les températures varient généralement entre -2°C et +2°C par rapport aux normes. +5,86°C, c'est énorme

	Résumé des données	Test statistique	Modélisation
Statistiques descriptives	Dans un lycée en fin d'année, les enseignants calculent les moyennes pour déterminer les passages et les redoublements. Pas de généralisation à faire, ces résultats ne concernent que leurs étudiants.	[[[Verdict pour un jugement ? Test sur le Sida ? Bref, un test qu'on n'a pas à reproduire]]]	Chuine travaille sur les années 1370 - 2003 et construit un modèle. Pas de généralisation à faire.
Statistiques inférentielles	[[[Audimat]]] Le résultat d'Harris n'est intéressant que parce qu'il s'applique à l'ensemble de la population	Oui ou non le téléphone portable augmente-t-il les chances de tumeur cérébrale ? La encore, cette information est importante si elle se généralise à l'ensemble de la population.	[[[Les modèles économiques ? Les modèles médicaux ?]]]

1.4 Statistiques, informatique... et papier-crayon

Il y a encore quelques années, les statistiques étaient synonymes de calculs longs et fastidieux. De nos jours, de puissants logiciels permettent d'obtenir tout ce que l'on souhaite en quelques clicks, même sur des ordinateurs familiaux.

Reste que si une véritable étude se fait par ordinateur, la compréhension des outils statistique passe par un calcul à la main (ou au moins à la calculatrice). En effet, si vous n'avez jamais fait vous-même le calcul, quand l'ordinateur vous aura effectué pour vous un test bien compliqué, vous ne saurez pas vraiment ce qui se cache derrière et il vous sera impossible d'interpretter correctement vos résultats.

Nous vous présenterons donc, pour chaque concept statistique abordé dans ce poly, deux exemples :

- Un exemple simple : construit artificiellement (les exemples issu d'études réelles sont toujours compliqués), ils permettront de faire le travail "à la main", pour une bonne compréhension. Malheureusement, ces exemples ne sont en général pas passionnant[6]. Pire, au vue des résultats, on a souvent envie de déclarer : "Tout ça pour ça ? C'était évident, il n'y avait pas besoin d'utiliser un outil statistique pour s'en rendre compte..." Attendez donc la suite !
- Un exemple réel : issu de la littérature scientifique (comme l'heure de Harris ou les vignes de Chuine), ils seront bien plus complets. En particulier, à la simple lecture des données, il est généralement impossible de tirer la moindre conclusion et les données seront trop denses pour être traitée à la main. Nous en viendrons à utiliser un logiciel statistique. Au passage, les données[7] de tels exemples, trop volumineux pour être dans un livre (et pour être recopier à la main) seront disponibles sur http ://christophe.genolini.free.fr/polystat.html

Pour les "calculs manuels", nous utiliserons soit une calculatrice, soit Open Office (l'équivalent gratuit d'Excel, téléchargeable sur http ://fr.openoffice.org/)

Le choix d'un logiciel statistique est plus délicat. Il existe globalement deux sortes de famille de logiciels statistiques : ceux à menu déroulant et ceux à ligne de commande. Les premiers sont plus simple à utiliser : pour calculer un indice statistique, il suffit de sélectionner les bonnes données et d'aller choisir un test dans un menu déroulant par un click souris. Ce système présente néanmoins un inconvénient majeur, la non reproductibilité de l'analyse. Une fois l'analyse terminée et le résultat sous vos yeux, vous ne pouvez plus voir le détail de votre analyse. Deux mois après, quand vous reprenez votre étude, vous n'avez plus accès aux choix que vous aviez effectué à l'époque. Enfin, si vous voulez modifier l'analyse ou ajouter des données, il vous faut tout recommencer...

En face, on trouve les logiciels à lignes de commande. Beaucoup moins conviviaux, il faut les programmer pour obtenir un résultat. Il y a donc une phase d'apprentissage du logiciel. Par contre, il est facile d'enregistrer le code (c'est un simple fichier), on peut donc le modifier, le ré exécuter sur des données différentes, vérifier les choix que l'on a fait...

[6] Enfin, on fera de notre mieux

[7] Si les auteurs ont donné leur accord ou si les données sont publiques. Sinon, nous "fabriquerons" des données semi-réelles permettant d'obtenir exactement les mêmes résultats que ceux que les auteurs ont publiés.

En pratique, nous utiliserons le logiciel "R". C'est un logiciel d'analyse statistique à ligne de commande. Il est gratuit et téléchargeable sur http ://www.r-project.rog. C'est, tout logiciel confondu (gratuit et payant), un des plus performant du moment[8]
Pour utiliser "R", il nous faudra donc écrire des programmes et les donner à "R". "R" exécutera ensuite les instructions. Pour écrire le programme en question, il faut un éditeur de texte. Sous Windows, l'éditeur de texte est notepad[9]. Mais nous vous recommandons plutôt d'utiliser Tinn-R, téléchargeable gratuitement sur http ://www.sciviews.org/Tinn-R/ et conçu pour fonctionner en conjonction avec R.

$$\text{VOUS} \overset{Tinn-R}{\longrightarrow} \text{Programme} \overset{R}{\longrightarrow} \text{Analyse}$$

[8]L'autre logiciel top niveau du moment est SAS. Lequel est le meilleur? La question ne se pose pas : "R" est plus performant dans certains domaines, SAS dans d'autres. Il est d'ailleurs amusant de constater que les statisticiens de haute voltige maîtrisent les deux outils et peuvent très bien faire une partie de leur analyse avec "R", préférer SAS pour une autre... Mais ne vous inquiétez pas, au niveau de ce poly, tous les logiciels statistiques sont équivalents.

[9]Word est un traitement de texte, pas un éditeur.

Chapitre 2

Les bases

2.1 De quoi parle-t-on ?

Avant d'entrer plus avant dans le vif du sujet, il nous faut définir un certain nombre de concept.

2.1.1 Définitions

Définition 2.1.1.a : Individu
Un **Individu** (ou **Sujet**, ou **Unité statistique**) est l'objet étudié.

Définition 2.1.1.b : Population
La **Population** est l'ensemble de tous les individus.

Dans l'étude de Harris, les individus sont des passants, la population est l'ensemble des 200 individus interrogés. Mais les individus ne sont pas toujours des hommes. Dans notre deuxième exemple, Chuine travaille sur les années. Dans ce cas, on préfère parler d'unité statistique. Les unités statistiques de Chuine sont les années, sa population est l'ensemble des années entre 1370 et 2003.

Formellement, le terme de population est ambiguë : en effet, Harris ne travaille pas sur la population qui l'intéresse (les américains) mais sur un échantillon de cette population. On devrait donc parler de population pour Chuine et d'échantillon pour Harris. En pratique, la population désigne à la fois le groupe sur lequel on travaille (et qui dans certains cas devrait s'appeler échantillon) et la population cible de l'étude. Cela ne prête généralement pas a confusion, mais il faut avoir à l'esprit que certaines populations n'en sont pas...

Maintenant que nous savons de quoi nous parlons, il nous faut préciser ce qui est mesuré :

Définition 2.1.1.c : Variable
Une **Variable** (ou **Caractère**) est-ce qui est étudié chez les individus (et qui a priori varie d'un individu à l'autre).

Dans ce poly, les variables seront notées en petites majuscules et entre crochets `[Variable]`

Définition 2.1.1.d : Modalités
Les **Modalités** (ou **Ensemble fondamental**) d'une variable sont toutes les valeurs que cette variable peut prendre.

Définition 2.1.1.e : Observation
Une **Observation** d'une variable est une valeur que cette variable prend effectivement.

Les observations sont notées entre parenthèses : `(Observation)`

Chez Harris, pour chaque individu, l'expérimentateur note le numéro de l'individu, la question qu'il a posé et la réponse obtenue, soit trois variables. La première variable `[Id]` est un identifiant unique qui

permet de distinguer les individus les uns des autres. Cet identifiant est généralement considéré à part et ne fait pas l'objet d'un traitement. La deuxième variable est [Question]. La troisième contient la [Reponse].
Les modalités de [Id] sont les nombres entiers de 1 à 200. Les modalités de [Question] sont (Heure+$) et ($seul). Les modalité de [Reponse] sont (Oui) et (Non).

Dans l'étude de Chuine, les variables sont [Annee], [Temperature] et [Date-de-recolte]. Les modalités de [Annee] sont toutes les années entre 1370 et 2003. Les modalités de [Temperature] sont toutes les températures comprises entres 0°C et 50°C. [Date-de-recolte] peut prendre toutes les valeurs allant du 1er janvier au 31 décembre.
Ici apparaît la différence entre les modalités et des observations. La [Date-de-recolte] peut *théoriquement* prendre pour valeur n'importe quel jour de l'année. En pratique, jamais une récolte n'a eu lieu le 2 février. (2 février) est donc une modalité (une valeur possible) mais pas une observation, la variable [Date-de-recolte] ne prenant jamais pour valeur (2 février).

Nous savons maintenant de quoi nous parlons et ce que nous mesurons. Nous nous retrouvons avec d'immenses tableaux remplis de données généralement trop vastes pour que l'esprit humain puisse les appréhender dans leur ensemble (comme le tableau1.2.1.a à partir duquel il est bien difficile de tirer un quelconque enseignement). D'où le besoin de "résumer" ces tableaux.

Définition 2.1.1.f : Indice statistique
Un **Indice statistique** (ou **Indice**) est une méthode de calcul permettant de résumer une grande quantité d'informations en une valeur unique.

En pratique, vous connaissez déjà beaucoup d'indices statistiques. Le plus célèbre d'entre eux est sans doute la moyenne : la moyenne générale d'un étudiant est un résumé de l'ensemble de ses notes. La moyenne générale d'une classe résume l'ensemble des notes de tous les élèves. Lors d'un conseil de classe, il est évidant que les enseignants ne peuvent pas donner toutes les notes de tous les élèves pour chacune des classes. Ils "compactent" donc des résultats individuels en un résultat global, la moyenne.
Harris utilise un autre indice encore plus élémentaire : il compte les "effectifs" de chaque modalité. Effectif d'une modalité est un indice résumant les données en les comptant, tout simplement. Nous verrons plus en détail tous ses indices (et bien d'autre) au chapitre prochain.

Comme nous le verrons, pour un même problème, il existe souvent plusieurs indices disponibles. Ce pose alors le problème du choix : lequel utiliser ? Pour aider à la décision, les statisticiens ont défini cinq qualités que tout indice respectable se doit d'avoir :

- **Cohérence définitoire** ou **Cohérance** : les statisticiens n'inventent pas des indices pour le plaisir ; chaque indice correspond à un problème pratique. "Ce vague concept auquel je suis en train de penser, se demande le chercheur, a-t-il vraiment du sens ? Peut on le définir précisément ?" C'est à ces deux questions que doit répondre un indice. Un bon indice a donc pour propriété de mesurer effectivement le concept pour lequel il a été crée. Cela à l'air trivial, mais ce n'est pas le cas de tous les indices.
  Malheureusement, vérifier cette propriété relève souvent plus de la philosophique que de la statistique. Un exemple assez classique est celui de l'intelligence. Le concept est indéniablement flou, mais le bon sens nous dit qu'il existe et qu'il serait utile de le mesurer. Les scientifiques ont donc mis sur pieds des indices, le plus célèbre étant le QI. Le QI mesure-t-il vraiment l'intelligence ? Difficile de dire parce que le concept d'intelligence est lui-même flou[1]. Reste que la difficulté de la tâche ne doit pas nous cacher l'importance de cette propriété : la cohérence est fondamentale, sans doute la plus importante propriété des indices.

- **Suffisance** : un indice est suffisant si son calcul utilise toutes les informations de l'échantillon. Cela peut semble aller de soi, mais en pratique, beaucoup d'indices n'utilisent qu'une partie des données. Considérons une population de 100 personnes sur laquelle nous calculons un indice suffisant (il utilise

[1] Pour aller un peu plus loin, il est envisageable d'inverser le processus et de définir le concept relativement à l'indice. Binet, cofondateur du QI, disait "L'intelligence ? C'est ce que mesure mon test". Derrière la boutade se cache pourtant une véritable idée et dans d'autres domaines, le pas a été franchi. Par exemple, pour évaluer le niveau global de l'élève, on a construit un indice appelé moyenne. Mais le processus s'est inversé, la moyenne est maintenant utilisé comme définition du niveau global...

donc les 100 personnes) et un indice non suffisant (qui par exemple n'utilise que deux personnes). Si la valeur d'une unique personne est modifiée :

  - L'indice suffisant est le mélange des informations de 100 personnes. Une unique valeur légèrement modifiée n'aura généralement que peu d'impact.
  - Par contre, le calcul de notre indice non suffisant se base sur deux personnes. Si la valeur modifiée touche l'une de ses deux personnes, l'indice peut changer du tout au tout. Sinon, il ne change pas du tout.

Les chercheurs n'aiment pas tellement l'idée que leurs résultats (de futures "vérités scientifique") puissent changer du tout au tout à cause d'une unique observation. La grande volatilité des indices non suffisants rend donc leurs résultats moins fort.

- **Absence de biais** : quand un chercheur a calculé un indice sur un échantillon, son désir le plus cher est de généraliser son résultat. Il a envie de déclarer : "Je viens de calculer un indice sur un échantillon, si je l'avais calculé sur toute la population, il prendrait la même valeur (à quelques décimales prêt bien sur) et donc mon résultat s'applique à tout le monde...". Affirmation bien audacieuse. Les statisticiens se sont penchés sur le problème et ont déterminé que dans certain cas, les indices prenaient effectivement la même valeur sur les échantillons et sur la population. Par contre, dans d'autre cas, les indices calculés sur un échantillon sous estime ou sur estime l'indice de la population correspondante. Ces indices sont dit **biaisés**. Un indice biaisé calculé sur un échantillon n'est pas généralisable à une population. D'où l'importance, pour les indices, d'être non biaisés.

- **Efficacité** : plus on travaille sur un échantillon de taille importante, plus augmente nos chances de trouver un indice proche de celui de la population. Ce principe est assez naturel : faites un sondage auprès de vos parents proches (échantillon de 5 ou 6 français) pour connaître le résultat des prochaines élections. Peu de chance que le résultat soit juste. Sondez maintenant tout votre quartier : vous aurez une approximation plus raisonnable. Et plus vous augmenterez la taille de votre échantillon, plus vous améliorerez la qualité de votre estimation. Cette règle générale (plus l'échantillon est important, plus l'estimation est bonne) s'applique à tout les indices non biaisé. Par contre, tous ne convergent pas à la même vitesse : certains indices donnent de bons résultats dès les petits échantillons, d'autres ne seront acceptables que sur des échantillons moyens, pour d'autres enfin, un échantillon important sera nécessaire. Un indice **efficace** est un indice qui donne de bon résultat dès les échantillons de petite taille. En pratique, cette notion est assez subjective et il est difficile de chiffrer l'efficacité d'un indice, mais on peut tout de même comparer les indices entre eux et savoir, parmi plusieurs, celui dont l'efficacité est la plus grande.

- **Robustesse** : dernière propriété essentielle, la robustesse. Nous le verrons très prochainement, les données collectés dans une expérience sont rarement "propres", elles contiennent des erreurs, des valeurs dites "aberrante". Une **Valeur aberrante** est une observation qui, à l'évidence, est fausse. Un cas typique, on demande à des patients de saisir leur date de naissance, à la place ils entrent la date du jour... Il est alors évident pour le chercheur que l'observation est fausse, la question a été mal interprêtée. Dans ce cas, c'est évident. Dans d'autres cas, c'est beaucoup moins facile à voir. Un indice est robuste s'il résiste à la présence de valeurs aberrantes, s'il ne se laisse pas ou peu influencer.

- **Polyvalence** : dans certain cas, il est souhaitable d'avoir la possibilité de calculer le même indice sur des données de type différent. Mais cette propriété n'est pas fondamentale et peut généralement être négligée.

- **Simplicité d'évaluation** : nous ne mentionnons cette dernière propriété que pour des raisons historiques. Pendant longtemps, la complexité de calcule de certains indices pouvait rendre leur usage délicat, on pouvait leur préférer des indices plus simple. Aujourd'hui, à l'aire des ordinateurs, ce critère n'a plus aucune raison d'être.

Naturellement, ces propriétés sont des indications de qualité et aucun indice ne les possède toutes complètement. En particulier, la suffisance et la robustesse sont intrinsèquement contradictoires : si un indice utilise toutes les données, alors il sera sensible aux valeurs aberrantes...

2.1.2 Nomenclature

Dans nos exemples, nous avons choisi pour nos variables des noms longs. Nous aurions pu choisir de les appeler `[X]` et `[Y]`. Mais `[Question-posee-par-Harris]` ou plus simplement `[Question]` est beaucoup plus explicite. Dans une petite étude statistique, ça n'a pas vraiment d'importance (encore que). Sur un travail conséquent, donner des noms explicites aux variables augmente grandement la lisibilité.

Dans ??, Connors et collaborateurs présentent une étude sur la pose d'un cathéter sur une artère cardiaque : certain patients reçoivent le cathéter, d'autre non. Connors note si un cathéter a été posé, il note aussi 59 autres variables (tableau 2.1.2.a).

1. ROWNAMES	2. QUANTAGE	3. QUALAGE	4. SEX	5. EDU
5. RACE	6. INCOME	7. PTID	8. CAT1	9. CAT2
11. SWANG1	12. DEATH	13. SADMDTE	14. DSCHDTE	15. DTHDTE
16. LSTCTDTE	17. CA	18. CARDIOHX	19. CHFHX	20. DEMENTHX
21. PSYCHHX	22. CHRPULHX	23. RENALHX	24. LIVERHX	25. IMMUNHX
26. GIBLEDHX	27. MALIGHX	28. AMIHX	29. RESP	30. CARD
31. NEURO	32. GASTR	33. RENAL	34. META	35. HEMA
36. SEPS	37. TRAUMA	38. ORTHO	39. APS1	40. SCOMA1
41. DAS2D3PC	42. ADLD3P	43. WTKILO1	44. TEMP1	45. HRT1
46. RESP1	47. MEANBP1	48. WBLC1	49. PAFI1	50. ALB1
51. HEMA1	52. BILI1	53. CREA1	54. SOD1	55. POT1
56. PACO21	57. PH1	58. URIN1	59. DNR1	60. TRANSHX

TAB. 2.1.2.a – Connors, liste des variables initiales

Ils remarquent que pour un patient `[44.TEMP1]` = `(33)`, `[2.QUANTAGE]` = `(19)` et `[28.AMIHX]` = `(Non)`. Que peuvent-ils en conclure ? Même le statisticien qui traite les données aura du mal à savoir sans se référer à la table des abréviations (il a du mal pendant qu'il est en train de traiter les données. Si par malheur, il termine puis est obligé de se replonger dans son étude trois mois plus tard, il ne se souvient clairement plus de rien). De même, quelle variable indique si chez un patient, un cathéter a été posé ? Impossible à dire.
En travaillant avec des variables dont le nom est explicite comme celle du tableau2.1.2.b, les résultats sont tout de suite beaucoup plus facilement interprétables :

1. IdPatient	2. Age	3. AgeParClasse	4. Sexe	5. Education
6. Race	7. Salaire	8. AssuranceMed	9. Pathologie1	10. Pathologie2
11. Catheterisation	12. Mort	13. DateAdmission	14. DateDechargeHopital	15. DateDeces
16. DateDerniereNouvelle	17. Cancer	18. InsuffisanceCardiaque	19. Prb_Cardiaques	20. Prb_Neurologiques
21. Prb_Psycologique	22. Prb_Pulmonaire	23. Prb_Renaux	24. Prb_Hepatic	25. Prb_Immunitaire
26. HemoragieGastroSup	27. TumeurSolide	28. InfarctusMyocarde	29. DiagRespiratoire	30. DiagCardiaque
31. DiagNeurologique	32. DiagGastrologique	33. DiagRenal	34. DiagMetabolique	35. DiagHematologique
36. DiagSepsis	37. DiagTraumatique	38. DiagOrthopedique	39. APACHE	40. GLASGOW
41. IndependanceDASI	42. IndependanceADL	43. Poids	44. Temperature	45. FrepCardiaque
46. FreqRespiratoire	47. TensionArterielle	48. HematoGlobulesBlancs	49. PressionArtO2Sonde	50. HematoAlbumine
51. HematoHematocrit	52. HematoBilirubin	53. HematoCreatine	54. HematoSodium	55. HematoPotassium
56. PressionArtCO2	57. HematoPH	58. Urine	59. NonResurection	60. Transfer

TAB. 2.1.2.b – Connors, liste des variables renommées

Si `[44.Température]` = `(33)`, clairement le patient a très froid. `[2.Age]` = `(19)`, il est tout jeune alors que `[28.InfarctusMyocarde]` = `(Non)` indique qu'il n'a pas eu d'infarctus du myocarde. Quant à savoir le traitement qui lui a été appliqué, il suffira probablement de regarder la variable `[11.Catheterisation]`

1. Quelle est la population considérée ?
2. Quels sont les individus ?
3. Quelle est, en plus de l'identifiant, la variables en jeu ?
4. Quelles sont les modalités de cette variable ?
5. Quelles est l'observation du sujet 2 ?
6. Donnez un exemple de modalité qui est également une observation.
7. Donnez un exemple de modalité qui n'est pas une observation.

2.2 Nature d'une variable

2.2.1 Principe de la classification

Les variables mesurent des choses extrêmement diverses. Cela va d'une couleur à une distance en passant par les appréciations notées sur des copies d'examen, l'état mental d'un patient, la consommation d'un biceps en oxygène...

En pratique, toutes les variables n'ont pas les mêmes propriétés. Par exemple, on peut faire la moyenne des températures de Chuine mais pas la moyenne des Oui / Non de Harris. Les variables sont donc classées en groupes selon leurs caractéristiques mathématiques. Trois propriétés intéressent particulièrement les statisticiens :

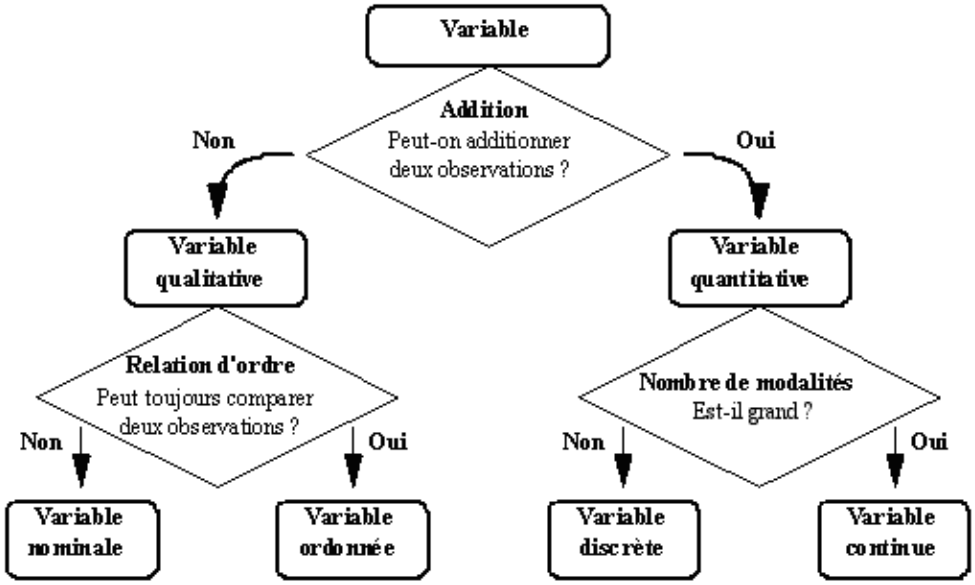
- **1. Arithmétique :** peut-on *additionner*[2] les observations entre elles ?
- **2a. Comparaison :** cela a-t-il du sens de dire qu'une observation *est plus grande* qu'une autre ?
- **2b. Continuité :** le nombre de modalités est-il *petit* ou *grand* ?

La réponse à ses trois questions détermine le type de notre variable.

- **1. Arithmétique :** Si on peut effectuer des additions[2], la variable est **quantitative**. Sinon elle est **qualitative**.
- **2a. Comparaison (ou relation d'ordre) :** Les variables quantitatives sont toujours comparables. Cette question ne se pose donc que pour les variables qualitatives. Une variable qualitative dont les modalités ne sont pas comparables est dite **nominale**. Si ses modalités sont comparables, elle est **ordonnée**
- **2b. Continuité (ou nombre de modalité) :** Cette question ne se pose que pour les variables quantitatives. Formellement, une variable dont le nombre de modalité est fini est dite **discrète**. Si le nombre de ses modalités est infini, elle est **continue**. En pratique, on fait une approximation : si une variable a un grand nombre de modalité, on la considère comme continue. Au final, une variable est discrète si elle a un petit nombre de modalité, elle est continue si elle a un grand nombre de modalité (ou une infinité de modalités).

La figure 2.2.1.a résume cette classification.

[2] Par addition, on entend addition et toutes les opérations arithmétiques habituelles : addition, soustraction, multiplication par un nombre, division, calcul de moyenne... Pour ne pas alourdir, nous ne parlerons que d'addition, mais gardez à l'esprit que toutes les autres opérations arithmétiques sont également possibles



Type	2a. Comparaison	1. Arithmétique	2b. Beaucoup de modalités
Variables qualitatives	?	Non	Non
- Variable nominale	Non	Non	Non
- Variable ordonnée	Oui	Non	Non
Variables quantitatives	Oui	Oui	?
- Variable discrète	Oui	Oui	Non
- Variable continue	Oui	Oui	Oui

FIG. 2.2.1.a – Nature des variables

Tout cela sera bien plus clair après quelques exemples.

2.2.2 Variables qualitatives

Les variables qualitatives contiennent des données qui s'évaluent avec des mots et non avec des nombres. Elles mesurent une qualité de l'individu et non une quantité : Le métier, le sexe, l'appartenance à un parti politique, le statut marital, le groupe sanguin, la mention au bac, le don ou non d'un dollar à Harris...

Les observations sont donc des mots et non des nombres, c'est pour ça qu'on ne peut pas les additionner.

Définition 2.2.2.a : Variable nominale
Une **Variable nominale** (ou **Variable qualitative pure**) est une variable dont on ne peut ni additionner, ni les ordonner les modalités.

La variable [CATÉGORIE SOCIO-PROFESSIONNELLE] est un exemple de variable nominale. Ses modalités sont (Agriculteurs exploitants), (Artisans commerçants / Chefs d'entreprises), (Cadres / Professions intellectuelles supérieures), (Professions intermédiaires), (Retraités), (Employés), (Ouvriers), (Autre). Clairement, additionner (Employés) et (Ouvrier), ou affirmer que (Employés) est plus grand que (Ouvrier) n'a pas de sens.

Définition 2.2.2.b : variable ordonnée
Une **Variable ordonnée** (ou **Variable semi-qualitative** est une variable dont on ne peut pas additionner les modalités, mais dont on peut les ordonner.

La variable [MENTION-AU-BAC] entre dans cette catégorie. (Passable) est moins bien que (Assez-bien), elle-même moins bien que (Bien) ou que (Très bien). Par contre, il n'est pas possible d'additionner (Assez-bien) et (Bien).

2.2.3 Variable quantitative (ou numérique)

C'est la variable la plus utilisé. Elle comprend toutes les mesures sur lesquelles il est possible de faire des opérations et des comparaisons : performances sportives, notes, nombre de membre de clubs, ...

Définition 2.2.3.a : Variable discrète
Une **Variable discrète** est une variable dont les modalités sont additionnables et peu nombreuses.

La variable [NOMBRE-D-ENFANT] peut prendre les valeurs 0, 1, 2, ... jusqu'à 10 ou 15 dans certains cas extrêmes, mais pas plus.

Par opposition aux variables discrètes, les variables continues sont des variables numériques pouvant prendre un grand nombre de valeur (soit un nombre infini, soit un nombre fini mais important). Par exemple, la variable [DISTANCE-DE-SAUT] peut prendre n'importe quelle valeur entre 0 et 10m. Elle peut prendre en particulier des valeurs comme 5,65m ou 5,66m mais aussi 5,658m ou 5,6583m... soit une infinité de valeur.
Naturellement, tout cela est très théorique puisque les moyens de mesure ne permettent pas une précision infinie. Dans la pratique, les variables continues sont généralement transformées en variables discrète : les modalités de [DISTANCE-DE-SAUT] seront les valeur 0,00m ; 0,01m ; 0,02m et ainsi de suite jusqu'à 9,99m ; 10,00m. Si un athlète saute 5,6583m (a supposé qu'on soit capable d'une telle précision), on arrondi au centième le plus proche.

Définition 2.2.3.b : Variable continue
Une **variable continue** est une variable dont les modalités sont additionnables et nombreuses.

2.2.4 Quelques pièges à éviter...

Nombre $\not\Rightarrow$ quantitative : la nature d'une variable se détermine exclusivement selon la classification que nous venons de voir. En particulier, si les modalités d'une variable sont des nombres, cela N'entraîne PAS automatiquement que la variable soit numérique. Par exemple, les modalités de la variable [DÉPARTEMENTS-FRANÇAIS] sont souvent des nombres. Pour autant, on ne peut pas faire la moyenne entre deux départements, ni prétendre que la Haute-Garonne est *supérieure* au Gers. D'où [DÉPARTEMENT-FRANÇAIS] est une variable qualitative nominale. De même, le [SEXE] est souvent noté en 0 pour femme, 1 pour homme. La encore, cette variable est nominale.

Ordonnée $\overset{?}{\Rightarrow}$ Quantitative : sous certaines conditions, il est possible de transformer une variable ordonnée en une variable numérique tout simplement en attribuant des valeurs aux variables ordonnées. Par exemple, si un prof note des copies en utilisant le système A, B, C, D et E (A pour très bien, E pour pas bien du tout), il peut estimer que A vaut 18, B vaut 14, C vaut 10, D vaut 6 et E vaut 2. Ce faisant, il vient de changer une variable ordonnée en numérique. Il peut donc ensuite faire un calcul de moyenne et toutes les statistiques liées aux variables numériques. Attention toutefois à la subjectivité d'une telle transformation : un autre prof aurait peut-être fait A=20, B=15, C=10, D=5 et E=0, ou encore A=20, B=16, C=12, D=8 et E=4. Les trois enseignants n'auraient pas obtenus la même moyenne. Autre problème, une variable numérique a des propriétés d'intervalle (la différence entre 5 et 10 est la même que celle entre 15 et 20) que n'a pas forcément la variable ordonnée (est-ce la même chose de passer de D à C que de B à A ? Rien n'est moins sur). Ce genre de transformation est donc à manipuler avec précaution. Pour plus de détail sur la numérisation des variables ordonnées et des échelles de mesures, je vous recommande l'excellent *Mesurer la subjectivité en santé* [Falissard01].

Quantitative $\overset{?}{\Rightarrow}$ Ordonnée : réciproquement, il peut arriver qu'une variable ordonnée ait plus de sens qu'une variable discrète ou continue. Par exemple, pour un urgentiste, l'age est un facteur pronostique important de décès. Les jeunes (généralement gravement accidenté) et les personnes agées (difficulté de récupération et fatigue générale) sont bien plus a risque que les autres. D'où, un urgentiste préfèrera la variable ordonné [TRANCHE-D-AGE] à la variable continue [AGE]. Le passage de la continue à l'ordonnée se fait simplement en définissant les classes d'age (par exemple (20-30), (30-40), (40-50), (50-60), (60-70), (70-80), (80-90) et (90-120)) puis en attribuant à chaque individu la classe dans laquelle son age le place.

Nominale $\overset{?}{\Leftrightarrow}$ Ordonnée : dans certain cas, la nature nominale ou ordonnée d'une variable qualitative peut relever plus du choix philosophique que mathématique : par exemple, les sociologues classent les métiers en Catégories Socio-professionnelles, variable généralement considérée comme nominale. Mais un sociologue peut parfaitement "décider" d'instaurer une relation d'ordre sur les CSP. Par exemple, il peut les ordonner selon le salaire moyen. Pour lui, `[Catégorie-Socio-Professionnelle]` sera donc une variable ordonnée. Un de ses collègues travaille sur le risque. Il ordonne donc les métiers en fonction des risques qu'ils font courir à ceux qui les exercent. Les deux sociologues n'auront pas classé les CSP dans le même ordre, mais tous deux considèreront pourtant cette variable comme ordonnée.

En pratique : même si ça n'est pas l'usage actuel, nous vous recommandons vivement de toujours utiliser les vraies modalités d'une variable. Par exemple, utilisez les modalités `(Homme)` et `(Femme)` plutôt que `(0)` et `(1)` pour la variable sexe. Même `(Haute-Garonne)` et `(Gers)` nous paraissent préférable a `(31)` et `(32)`. Bien sûr, c'est plus long à taper[3]. Mais vous y gagnerez beaucoup, à deux points de vue : tout d'abord, vos résultats seront beaucoup plus lisibles. "La `[CSP]` majoritaire est la `(4)`" est bien moins parlant que "La `[Categorie Socio-Professionnelle]` majoritaire est `(Profession libérale)`". Ensuite, et c'est un point très important, les logiciels statistiques actuels sont capables de déterminer tout seul le type d'une variable, puis de choisir l'analyse statistique correspondante. Si vous codez `[Sexe]` en 0 / 1, le logiciel identifiera votre variable comme numérique et lui appliquera automatiquement[4] l'analyse statistique des variables numériques... ce qui dans notre cas est clairement faux !

2.2.5 Récapitulatif

Nous savons maintenant ce qu'est une population, des individus, des variables et leurs modalités. La première étape d'une analyse sera de dresser un tableau résumant ces différents paramètres sous forme synthétique.
Les variables de l'étude de Harris sont présentées table 2.2.5.a.

Variable	Nature	Modalités
Question	Nominale	Heure + $; $ seul
Reponse	Nominale	Oui ; Non

TAB. 2.2.5.a – Harris, liste des variables

Les variables de l'étude de Chuine sont présentées table 2.2.5.b.

Variable	Nature	Modalités
Annee	Continue	[1200 ; 1950]
Temperature	Continue	[0 ; 60]
JourRecolte	Continue	[1 ; 365]

TAB. 2.2.5.b – Chuine, liste des variables

2.2.6 Exercice

Dans chacun des cas,

1. Dans un questionnaire de satisfaction, la direction d'un parc d'attraction demande à ses clients leur opinion sur la climatisation des salles de cinéma. La `[température]` est-elle Trop froide ? Correcte ? Trop chaude ?
   (a) Donnez deux exemples de modalités possibles.
   (b) Vérifiez les propriétés mathématiques de la variable.
   (c) En conclusion, déterminez le type de variable.

[3] Encore que, avec l'utilisation des ordinateurs, vous pouvez très bien saisir des nombres au clavier, puis les transformer en Nom juste avant l'analyse
[4] C'est-à-dire si vous n'y prenez pas garde. Mais n'est-ce pas courir un risque bien inutile ?

2. Dans une étude sur les illusions visuelles, un expérimentateur demande aux patients de préciser la couleur d'un objet. Il mesure la variable `[Couleur]`. Reprenez les trois points (a), (b) et (c) si dessus et déterminez le type de variable
3. [[[variable continue. I comme icare ?]]]

Chapitre 3

Analyse univariée

3.1 Généralité

3.1.1 Le principe

L'analyse univarié consiste à étudier les variables indépendamment, une à une. Pour chacune d'entre elles, nous calculerons trois types d'indices.

- **1. Effectifs :** la première étape est de rassembler les données. Nous avons déjà vu, dans l'introduction avec Harris, que compter les gens qui ont eu le même comportement donne une première idée de ce qui se passe. Cette étape vient compléter celle que nous venons de voir dans le chapitre précédent, la **définition des variables** (préciser les variables sur lesquelles nous allons travailler et leurs modalités possibles.)
- **2. Centralité :** les indices de centralité permettent de définir approximativement le "centre" des observations. Différentes définitions de centre ont donné naissance à différents indices parmi lesquels on trouve la moyenne (grande star de la statistique), le mode et la médiane.
- **3. Dispersion :** les indices de dispersion indiquent le regroupement (ou non) des observations autour de la valeur centrale. Là encore, différentes définitions de "groupé" conduisent à différents indices. Le plus célèbre est l'écart type.

Une fois ces indices calculés, on termine en les représentant graphiquement. Cette partie, souvent considérée comme accessoire par le débutant, est pourtant fondamentale.[1]. Il est même certaine situation ou les tests statistiques ne peuvent pas trancher (tranchent mal) et où le recours à une représentation graphique est la seule solution.

3.1.2 Exemple

Avant d'entrer dans le détail de l'analyse univariée, un petit exemple va permettre de voir son utilité. L'exemple est tiré de situation réelle, à peine modifié pour les besoins de la cause.
Le jeudi après midi, j'enseigne les statistiques. Mes étudiants sont, comme tous les étudiants, des jeunes gens très sérieux mais vivant dans un milieu naturellement riche en distraction et peu propice à un travail régulier, régularité pourtant nécessaire à la bonne marche de leurs études. Pour les aider, j'ai donc instauré un contrôle continu hebdomadaire. Chaque semaine, je me retrouve donc avec un paquet de copies anonymes (le desanonymage se fait en fin de semestre en une seule fois, sous contrôle des délégués étudiants) que je note. Après les 5 premières semaines, les résultats sont les suivants : Semaine 1 : 8, 6, 3, 6, 6, 8, 5, 7, 9, 7, 6, 8, 7, 9, 8, 9. Semaine 2 : 12, 15, 13, 15, 15, 18, 13, 14, 13, 14, 12, 16, 14, 12, 13, 12. Semaine 3 : 3, 10, 11, 7, 15, 10, 18, 7, 8, 12, 15, 16, 8, 14. Semaine 4 : 3, 11, 10, 3, 4, 3, 12, 5, 18, 5, 11, 4, 6, 12, 8, 13. Semaine 5 : 18, 10, 11, 13, 17, 18, 9, 16, 3, 16, 10, 17, 15, 9, 18, 8.
Au passage, les copies étant anonymes, l'ordre des notes ne correspond pas à l'ordre des étudiants.

A partir de là, que dire ?

3.1.2.1 Effectifs

Pour commencer, précisons les objets sur lesquels nous allons travailler :

[1] Les autres, faites comme vous voulez, mais pour mes étudiants, la représentation graphique est O-BLI-GA-TOIRE !

- Un individu est ici une copie
- Pour chaque individu, nous mesurons deux variables :
  - [Semaine] peut prendre les valeurs 1, 2, 3, 4 et 5.
  - [Note] peut prendre toutes les valeurs entières entre 0 et 20.

Ensuite, nous comptons. Nous pouvons assembler les individus (copies) de plusieurs manières. Ici, celle qui nous semble la plus naturelle est de grouper les copies par semaine (comme nous le verrons plus tard, il y aurait bien d'autre manière de rassembler les données). Ensuite, nous comptons le nombre de copie de chaque semaine :

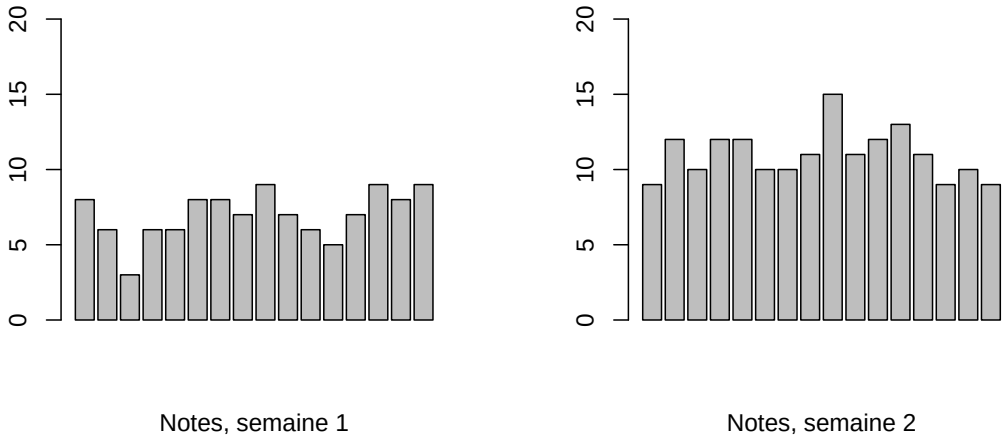
Semaine	Nombre
1	16
2	16
3	14
4	16
5	16

Tab. 3.1.2.a – Contrôle continu, effectifs par semaine

Première information, il y avait deux absents la troisième semaine.

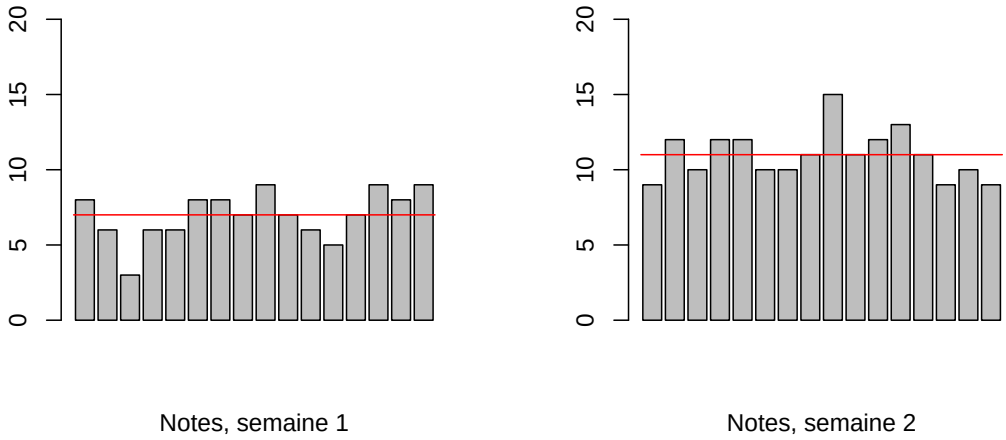
3.1.2.2 Centralité

Pour bien comprendre l'intérêt des indices de centralité, représentons graphiquement les notes obtenues la première et la deuxième semaine. Chaque note est représentée par une colonne (meilleure est la note, plus haute est la colonne) :



Tab. 3.1.2.b – Contrôle continu, comparaison des semaines 1 et 2

Que nous apportent ces graphiques ? De manière informelle, ils nous donnent l'impression que la deuxième semaine, le contrôle a été bien mieux réussi que la première. Cette impression est-elle justifiée ? et si oui, comment la formaliser ? Autrement dit, pourrait-on trouver un indice qui résume d'un seul coup toutes les notes d'une semaine et qui rende compte du fait que les notes de la deuxième semaine sont meilleures que celle de la première ? Plusieurs réponses sont possibles. Les statisticiens nous proposent en particulier d'utiliser un indice de centralité (pour l'instant, peut importe lequel et peut importe sa méthode de calcul), c'est-à-dire un indice dont la valeur se situe le plus possible au "milieu" de toutes les notes. Cette valeur est représentée graphiquement sur la figure suivante par la ligne rouge :

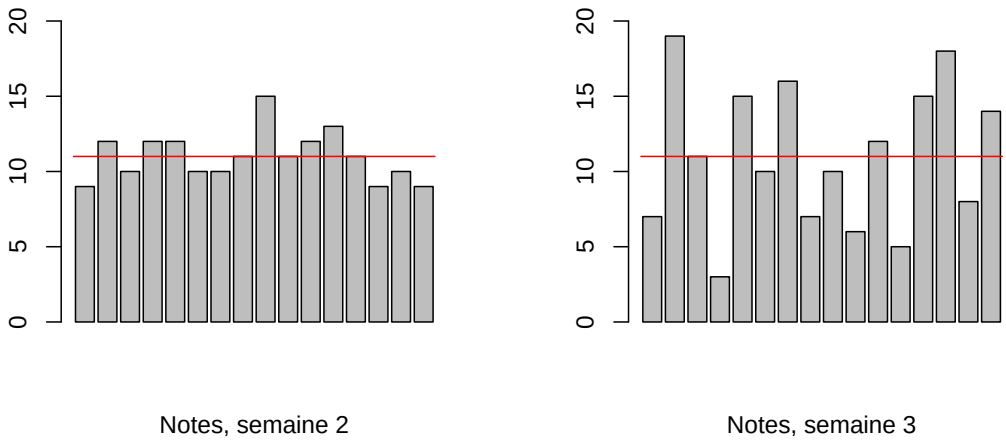


TAB. 3.1.2.c – Contrôle continu, semaines 1 et 2 avec indice de centralité

Graphiquement, on constate bien que les notes "gravitent" autour de la ligne. La valeur de l'indice, correspondant à la hauteur de la ligne, est donc une valeur qui représente raisonnablement bien l'ensemble des notes[2]. C'est un indice de centralité. Il vaut 7 pour la première semaine et 13 pour la seconde. Globalement, il nous dit "La première semaine, les notes gravitaient autour de 7, le deuxième semaine, elles tournent autour de 13".

3.1.2.3 Dispersion

Pour comprendre l'intérêt de la dispersion, représentons graphiquement les notes obtenues la deuxième et la troisième semaine (graphique sur lequel nous plaçons d'ors et déjà l'indice de centralité) :

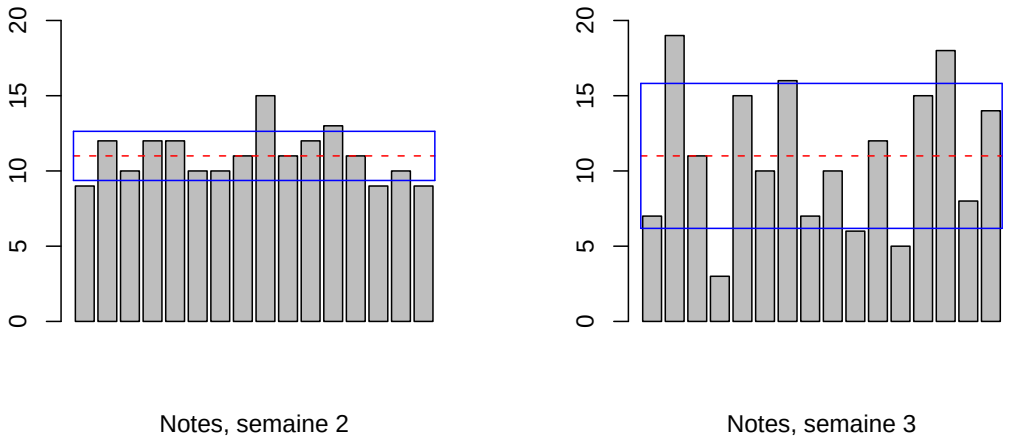


TAB. 3.1.2.d – Contrôle continu, comparaison des semaines 2 et 3

L'indice de centralité prend la même valeur sur les deux semaines. Pourtant, les graphiques nous donnent

[2]Vous n'êtes pas d'accord ? Vous avez parfaitement le droit et je le reconnais, cette affirmation est très subjective. Malheureusement (pour ceux qui ne sont pas d'accord) et heureusement (pour tous les autres), elle est aujourd'hui acceptée par pratiquement tout le monde et cet indice se retrouve partout

clairement l'impression que les deux contrôles sont différents. Plus précisément, la semaine deux présente des résultats plus homogènes que la semaine trois : en semaine deux, toutes les notes sont assez proches de la ligne rouge (toutes les notes sont proches de l'indice de centralité). En semaine trois, certaines notes sont proches mais d'autres sont éloignées de cette fameuse ligne rouge. Formaliser cette uniformité ou cette diversité est le but des indices de dispersion. Un indice de dispersion est un indice qui prendra une grande valeur si les notes sont dispersées et une petite valeur si les notes sont regroupées autour de l'indice central. Le calcul et la représentation graphique des indices de dispersion est un peu plus compliqué. Ici, nous représentons la dispersion par une boite bleue : une boite étroite dénote une faible dispersion, une boite large représente une grande dispersion.

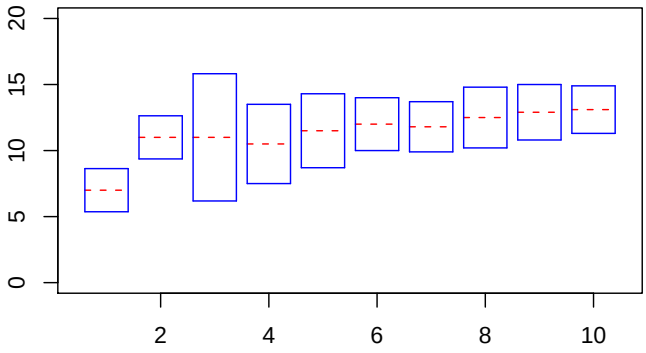


TAB. 3.1.2.e – Contrôle continu, semaines 2 et 3 avec indice de dispersion

Globalement, l'indice de dispersion nous dit "En semaine 2, le niveau des étudiants était plutôt homogène. En semaine trois, il est plutôt disparate."

3.1.2.4 Petit bilan

Il est maintenant temps de conclure avec cet exemple. A la réunion de fin d'année, lorsque l'équipe pédagogique se retrouve pour faire le bilan des 12 semaines d'enseignement, plutôt que d'apporter dans mon sac les $16 \times 12 = 192$ copie moins les quelques absences, je présenterais à mes collègues ce petit graphique :



TAB. 3.1.2.f – Contrôle continu, bilan

Petit graphique qui me permettra de déclarer : “A part la semaine 2, ce groupe TD est plutôt homogène (tous les carrés bleus sont petits) et a régulièrement progressé tout au long du semestre (au fur et à mesure que les semaines passent, les petits traits rouge sont de plus en plus haut)” !

Il est maintenant temps de détailler un peu plus les différents indices évoqués.

3.2 Effectif et distribution

3.2.1 Les données du problème

En 2003, lors de sa participation à la coupe du Comminges de rugby, le Racing Club de Villeneuve de Rivière (Haute Garonne) ne s'est pas trop mal débrouillé. Bravo à eux. Pour chaque match, l'entraîneur (un fin statisticien) a noté le numéro du match, le résultat final[3], le nombre de points marqués par son équipe et le petit commentaire laconique qu'il ne manque jamais de lâcher lorsque son équipe rentre dans les vestiaires...

[Match]	[Resultat]	[Points]	[Cartons]	[Meteo]	[Commentaire]
1	Perdu	0	0	Soleil	Mauvais
2	Gagné (bonus)	32	1	Soleil	Excellent
3	Gagné	25	0	Pluie	Mauvais
4	Nul	11	0	Soleil	Moyen
5	Gagné (bonus)	30	1	Soleil	
6	Gagné (bonus)	42	0	Pluie	
7	Perdu	6	0	Nuageux	Ultra nul
8	Perdu	22	0	Soleil	Mauvais
9	Gagné	21	1	Nuageux	Mauvais
10	Gagné (bonus)	31	0	Soleil	
11	Perdu	17	2	Soleil	Mauvais
12	Nul	0	3	Neige	Ultra nul
13	Perdu	15	2	Nuageux	Ultra nul
14	Gagné	27	0	Soleil	Moyen
15	Perdu	18	1	Pluie	Bon
16	Gagné (bonus)	36	0	Soleil	Excellent

Tab. 3.2.1.a – RCV, données brutes

3.2.2 Définition des variables

Avant de nous lancer dans des comptages, nous devons préciser les variables sur lesquelles nous allons travailler :

[3] En Rugby, une équipe qui gagne en marquant plus de 30 points reçoit en plus un point de bonification. Il existe donc deux types de victoire, la *victoire simple* et la *victoire avec bonus*

Variable	Nature	Modalités
[Resultat]	Qualitative ordonnée	Perdu Nul Gagné Gagné (bonus)
[Points]	Quantitative continue	$[0\,;+\infty]$
[Carton]	Quantitative discrète	0 ; 1 ; 2 ; ...
[Meteo]	Qualitative nominale	Soleil Pluie Nuageux Neige
[Commentaire]	Qualitative ordonnée	Très mauvais Mauvais Moyen Bon Excellent

Tab. 3.2.2.a – RCV, liste des variables

[Match] est un identifiant, il n'est donc pas considéré comme une variable.

3.2.3 Effectif

La liste des résultats telle qu'elle est présenté table 3.2.1.a est appelée "liste brute". On l'a déjà vu dans l'introduction, cette présentation n'est pas très astucieuse et on préfèrera donc, plutôt qu'une liste brute, regrouper ensemble les matchs dont les observations sont identiques, par exemple pour la variable résultats, les "Perdu" avec les "Perdu", les "Gagné" avec les "Gagné"... Au final, il y a eu 16 matchs répartis de la manière suivante :

[Resultat]	Nombre d'individu par groupe
Perdu	6
Nul	2
Gagné	3
Gagné (bonus)	5
Total	16

Tab. 3.2.3.a – RCV, effectifs de [Resultat]

Pour obtenir ce deuxième tableau, il a suffit de compter : le nombre total d'individu (ici, un individu est un match) et le nombre d'individu dont les observations (c'est à dire `(Perdu)`, `(Nul)`, `(Gagné)` et `(Gagné (bonus))`) sont identiques entre elles. Dans le jargon du statisticien, ces deux nombres s'appellent *effectif* et *effectif d'une modalité.*

Définition 3.2.3.a : Effectif
L'**Effectif d'une population** est le nombre total d'individu que contient la population. On le note généralement n.
L'**Effectif d'une modalité** est le nombre d'individu pour lesquels la variable prend la valeur de la modalité. On le note n_i.

Dans notre exemple, l'effectif de la population est 16, l'effectif de la modalité `(Gagné)` est 3.

Attention Piège : Il est important de faire la distinction entre la variable et ses effectifs. Dans notre exemple, la variable est [Resultats]. Le *nombre d'individu du groupe* N'est PAS une variable, c'est l'effectif d'une variable... La table 3.2.1.a présente les observations de la variable, la table 3.2.3.a présente les effectifs de la variable.

Fort des effectifs, nous pouvons compléter notre tableau :

Variable	Nature	Modalités : Effectifs
[Resultat]	Qualitative ordonnée	Perdu : 6 Nul : 2 Gagné : 3 Gagné (bonus) : 5

Tab. 3.2.3.b – RCV, nature et effectifs de [Resultat]

3.2.4 Distribution

Le concept de distribution découle directement de celui d'effectif ; une distribution est simplement la liste des couples *(Observations, Effectifs de l'observation)*.

Définition 3.2.4.a : distribution
Etant donnée une variable dont les modalités sont x_1, x_2, ... x_i, la *distribution* de la variable est la suite des couples {(x_i, Effectif de x_i}.

$$D = \{(x_1, n_1), (x_2, n_2), ...(x_i, n_i)\}$$

Dans notre exemple, la distribution de [Resultat] est le contenu de la colonne **Modalités : Effectifs** dans la table 3.2.3.b, à savoir {(Perdu,6) ; (Nul,2) ; (Gagné,3) ; (Gagné (bonus),5)}.

Ici apparaît la première différence entre les variables nominales et les ordonnées. Dans une nominale, la place des modalités dans la distribution est libre. Pour une ordonné, l'agencement des modalités se fait par ordre croissant. Chez Harris, la distribution de la variable [Reponse] (qui, nous le rappelons, est une variable nominale donc non ordonnée) peut être {(Oui,52) ; (Non,148)} ou {(Non,148) ; (Oui,52)}. Par contre, la distribution de [Resultat] de la coupe du Comminges N'aurait PAS pu être {(Gagné,3) ; (Perdu,5) ; (Gagné (bonus),4)) ; (Nul,2)}. Cela deviendra particulièrement important lorsque l'on passera à la représentation graphique de la distribution.

Variable	Modalités : Effectifs
[Reponse]	Oui : 52 Non : 148

Variable	Modalités : Effectifs
[Reponse]	Non : 148 Oui : 52

Tab. 3.2.4.a – Variable nominale : l'ordre est libre

Variable	Modalités : Effectifs
[Resultat]	Perdu : 6 Nul : 2 Gagné : 3 Gagné (bonus) : 5

Variable	Modalités : Effectifs
[Resultat]	Gagné : 3 Perdu : 6 Gagné (bonus) : 5 Nul : 2

Tab. 3.2.4.b – Variable ordonnée : l'ordre est imposé

[[[Distribution conditionnelle -> Bi-variée ou ici ?]]]

3.2.5 Fréquence

L'effectif est un nombre absolu, c'est à dire qu'il est indépendant des autres résultats. A ce titre, il peut ne pas être très informatif. Par exemple, un joueur de l'équipe de Villeneuve de Rivière déclare avoir perdu deux matchs cette saison. Que peut-on dire de ce joueur ? Rien, il manque une information, à savoir le nombre de match joués : s'il a perdu 2 matchs et n'en a joué que 2, c'est plutôt décevant. S'il a joué 11 matchs, c'est plutôt bien.[[[exemple de la recette de cuisine ? 300 g de beurre dans un gâteau pour 4, c'est beaucoup, dans un gâteau pour 15, c'est raisonnable]]]
D'où l'utilisation d'un indice qui prend en compte le nombre total d'observation : la fréquence.

Définition 3.2.5.a : Fréquence
La **Fréquence** d'une modalité x_i est le nombre d'individus n_i pour lesquels la variable prend la valeur x_i *divisé* par le nombre total d'individu. Le plus souvent, la fréquence est donnée en pourcentage. Elle est notée f_i.

$$f_i = \frac{n_i}{n}$$
$$f_i = \frac{n_i}{n} \times 100\%$$

La fréquence d'une population n'est pas considérée puisque toujours égale à 100%. Sur notre exemple, la fréquences de `(Perdu)` est : $f_{(\texttt{Perdu})} = 6/16 = 0,375 = 37,5\%$
On peut donc compléter notre tableau de présentation des variables :

Variable	Nature	Modalités : Effectifs (pourcentage)
[RESULTAT]	Qualitative ordonnée	Perdu : 6 (37,5%) Nul : 2 (12,5%) Gagné : 3 (18,75%) Gagné (bonus) : 5 (31,25%)

TAB. 3.2.5.a – RCV, nature, effectifs et fréquence de [RESULTAT]

En pratique, à chaque fois qu'on présente un pourcentage, il est indispensable de préciser la population à laquelle il s'applique (ou de donner le chiffre absolu associé, ce qui revient au même). Par exemple, on peut facilement trouver dans les journaux ou à la télévision des pseudo sondages du genre "48% des jeunes interrogés affirment avoir déjà consommé de la drogue". Quel crédit peut-on apporter à ce genre d'affirmation ? Cela dépend du nombre de sondés. Si le pourcentage est obtenu après avoir contacté 1000 jeunes (c'est le cas des statistiques présenté par exemple par l'OFDT[4]), il est fiable. S'il est obtenu auprès de 100 personnes (c'est généralement le cas dans les émissions de télé grand public du style "Attention à la marche"), il a très peu de valeur... Plus de détails sur l'importance des chiffres absolus et relatifs dans le chapitre "Comment tricher en statistique ?"[[[Sur 100 personnes, 78% acceptent de venir par téléphone, 50% viennent. Combien sont venus ? + Exemple des collèges]]].

3.2.6 Données manquantes

L'entraîneur de l'équipe de Villeneuve de Rivière est (selon les dires de son auxiliaire) un homme chargé de lourdes responsabilités qui ne lui laissent pas toujours le loisir d'assister aux matchs de son équipe. Cette saison, il en a raté 3 (les matchs 5,6 et 10). L'auxiliaire a scrupuleusement noté toutes les données qu'il pouvait, mais a été bien incapable de deviner le commentaire que l'entraîneur aurait lâché. La variable [COMMENTAIRE] présente donc ce qu'on appelle des *données manquantes*.

Les données manquantes font malheureusement parti du quotidien du statisticien. On en trouve dans toutes les études et il n'existe pas de technique permettant de bien les gérer. Du coup, comme on ne sait pas quoi en faire, elles sont taboues et rares sont les livres de statistique qui en parlent. Nombre d'auteurs se contentent donc purement et simplement de les ignorer. Pourtant, elles peuvent être responsables de biais importants.
Retour à Villeneuve de Rivière. Que dire sur les commentaires de l'entraîneur ? Les effectifs de la variable sont les suivants :

Variable	Nature	Effectifs
[COMMENTAIRE]	Ordonnée	Ultra nul : 2(15,38%) Mauvais : 4(30,77%) Moyen : 4(30,77%) Bon : 1(7,69%) Excellent : 2(15,38%)

TAB. 3.2.6.a – RCV, effectifs et fréquence de [COMMENTAIRE]

`(Moyen)`, `(Bon)` et `(Excellent)` ont la majorité avec 53,85%. Mais il y a trois données manquantes. Après enquête, nous réussissons à trouver l'entraîneur et à lui demander pourquoi certains de ses commentaires

[4] Observatoire Français des Drogues et des Toxicomanies

sont manquants. Sa réponse ne laisse aucun doute : "Parce qu'ils jouaient tellement comme des fada (sic), que si j'étais resté jusqu'au bout du match, j'en aurai pris un pour assommer l'autre (re-sic) !". Cela change complètement la nature des données : elles ne sont pas manquantes par hasard, elles sont manquantes les jours où, selon l'entraîneur, l'équipe jouait particulièrement mal. Si l'entraîneur avait eu la force de rester jusqu'à la fin, il aurait probablement marqué `(Ultra nul)`... Et le bilan serait le suivant[5] :

Variable	Nature	Effectifs
[COMMENTAIRE]	Ordonnée	Ultra nul : 5(31,25%) Mauvais : 4(25,00%) Moyen : 4(25,00%) Bon : 1(6,25%) Excellent : 2(12,50%)

TAB. 3.2.6.b – RCV, effectifs et fréquence de [COMMENTAIRE] *après correction*

`(Moyen)`, `(Bon)` et `(Excellent)` n'ont plus la majorité et le pourcentage d'Ultra nul a plus que doublé !

Cet exemple a l'air anecdotique, malheureusement, il ne l'est pas : dans une étude sur l'alcoolémie et les accidents de la route avec dommages corporels [Reynaud02], Reynaud constatent que dans 30% des cas, le taux l'alcool n'est pas noté (alors que c'est obligatoire). Après enquête, il s'avère qu'il n'est pas relevé quand le conducteur est physiquement inaccessible, ce qui dénote tout de même un accident particulièrement grave... Cela signifie que les données manquantes concernent principalement les accidents les plus graves... qui sont généralement liés à une forte alcoolémie. Ne pas considérer les données manquantes entraîne donc une sous estimation de nombre d'accidents dans lesquels l'alcool est impliqué.

Ce constat est généralisable : les parents maltraitant leurs enfants ne donnerons pas un nombre de coup quotidien ; les alcooliques vont omettre de répondre à la question sur la quantité d'alcool consommé ; plus généralement, la honte pousse le patient à ne pas répondre alors que c'est justement lui qui intéresse l'expérimentateur. De même, dans les questionnaires de satisfaction, seul les mécontents se manifestent. Les contents prennent rarement la peine de remplir et de renvoyer, donc les données des satisfaits sont manquantes. Inversement, quand des patients reçoivent un nouveau type de traitement, les patients satisfaits continuent à venir, les insatisfaits arrêtent et changent d'hôpital... Ceux qui refusent le système social ne répondront pas à un sondage... Comme vous pouvez le constater, les valeurs manquantes sont rarement dues au hasard et posent un vrai problème.

Que faire ? A notre niveau[6], trois choses sont possibles :

- En première lieu, l'expérimentateur doit tout faire pour avoir des données aussi complètes que possible : dans les cas cités ci-dessus, plutôt que de donner un questionnaire à un alcoolique, l'expérimentateur pourra essayer de faire passer des entretiens. Bien mené, si le patient est mis en confiance, ils donnent beaucoup moins de données manquantes. Ou encore, il pourra relancer les gens qui ont abandonné l'étude, les contacter pour qu'ils renvoient le questionnaire, bref, tout faire pour avoir des données les plus complètes possibles.
- La deuxième chose à faire relève plus de la franchise et de l'honnêteté que de la statistique pure : il faut systématiquement signaler les données manquantes. Cela n'a l'air de rien, mais c'est capital. En effet, la finalité des statistiques est d'établir des règles ("Il y a 95%de chance que..."), comme nous l'avons vu dans l'introduction). Or, si vous travaillez sur des données incomplètes, la règle que vous êtes en train d'établir est peut-être biaisée. Il vous faut donc le signaler : "J'ai trouvé telle règle mais je travaillais sur des données complètes à 70% seulement".
- Enfin, la majorité des études sont faites dans le but de montrer quelque chose. On peut alors remplacer les données manquantes par le résultat qui est *le plus défavorable* a ce que l'on veut

[5] Nous présentons ce bilan corrigé à titre pédagogique, pour mettre en valeur l'impact que les données manquantes peuvent avoir. En pratique, on ne peut pas "improviser" la valeur des données manquantes. Même si l'entraîneur nous promet qu'il aurait mis `(Ultra nul)`, peut-être un essai dans les cinq dernières minutes l'aurait fait changer d'avis. Donc, pas de correction possible, la variable [COMMENTAIRE] désigne le commentaire fin de match, pas un commentaire fait en fin de saison sans avoir vu l'intégralité du match.

[6] Pour information et pour plus tard, il existe tout de même des techniques plus ou moins valables permettant de diminuer l'impact des données manquantes. En particulier, les imputations ou imputations multiples donnent de bons résultats ; LOCF est a bannir.

montrer. Un statisticien souhaitant montrer que l'entraîneur du Racing Club Villenois est un vieil acariâtre (donc beaucoup d'appréciations négatives) devra remplacer les valeurs manquantes pas des `(Excellent)` ; à l'inverse, pour montrer qu'il fait beaucoup d'éloges à son équipe il devra les remplacer par des `(Ultra nul)`. On comprend dès lors toute l'importance de ne pas avoir de données manquantes !

En pratique, les étapes un et deux sont obligatoires. La troisième ajoute plus de crédibilité à l'étude, mais risque aussi de la rendre inintéressante. Dans tous les cas, on ajoute à la présentation des données une colonne qui précise le nombre de valeurs manquantes.

Variable	Nature	Effectifs	Manquantes
[COMMENTAIRE]	Ordonnée	Ultra nul : 2(15,38%) Mauvais : 4(30,77%) Moyen : 4(30,77%) Bon : 1(7,69%) Excellent : 2(15,38%)	3

TAB. 3.2.6.c – RCV, effectifs, fréquence et données manquantes de [COMMENTAIRE]

3.2.7 Valeurs aberrantes

Pire que les données manquantes, on trouve les valeurs aberrantes. Une valeur aberrante est une donnée qui, à l'évidence, est fausse. Si un médecin prend votre température et trouve 15°, il ne va pas spécialement s'inquiéter, il va changer de thermomètre. Pourquoi ? Parce qu'à l'évidence, votre température ne peut pas être de 15°. Une telle donnée indique donc une erreur de mesure. Si le Racing Club de Villeneuve marque 617 poins sur un match, vous allez à l'évidence considérer qu'il y a eu une erreur de notation lorsque l'entraîneur a reporté le score en fin de match. Si sur un questionnaire vous demandez la date de naissance et qu'on vous donne la date du jour, vous saurez qu'à l'évidence, la personne aura mal lu la question.

Malheureusement, toutes les valeurs aberrantes ne sont pas aussi simple à détecter. Dans**??**, Connors reporte un patient dont le taux l'albumine[7] est 29. Valeur aberrante ou non ? Peut-être qu'un médecin s'arracherait les cheveux en disant que non, personne ne peut vivre avec un taux d'albumine de 29 et qu'à l'évidence, c'est une valeur aberrante. Mais l'évidence pour lui n'est pas l'évidence pour le statisticien... Dans ces cas là, quand le statisticien ne sait pas, il peut toujours demander à un expert. Mais une valeur aberrante peut avoir toutes les apparences d'une valeur innocente pour le commun des mortels, ce qui fait que le statisticien n'aura même pas l'idée de contacter un spécialiste. Exemple, un groupe de sportives courent un 200m et annoncent {22,03s ; 23,56s ; 24,13s ; 248,2}. Le statisticien va tout de suite repérer la valeur aberrante : 248,2s pour un 200m, c'est à l'évidence impossible. Mais il passera sans doute à côté du 22,13s, situé à plus d'une seconde sous le record de France...

Pire, dans certain cas, les spécialistes eux même ne peuvent pas trancher : Toujours dans**??**, Connors reporte des patients dont les températures sont les suivantes : {27 ; 28 ; 28 ; 28,3 ; 29,5 ; 30,2 ; 30,9 ; 31 ; 31,1 ; 31,3 ; 31,4 ; 31,7 ; ...} A priori, personne ne peut survivre a 27°. Mais à 35°, c'est possible. Dès lors, quelles sont les valeurs aberrantes ? Un spécialiste dira 32°. Un autre affirmera qu'aux urgences, on voit des choses vraiment extrêmes et il placera la barre à 30°. Un autre aura une autre idée. D'où un problème réel : déterminer ce qui est aberrant n'est pas toujours possible.
En pratique, comment faire ?

- Les données qui sont trivialement aberrantes peuvent être purement et simplement supprimées. Elles sont alors considérées comme des données manquantes. Il faut naturellement préciser qu'elles ont été supprimées, et justifier leur suppression.
- Pour les données sur lesquelles il y a un doute, les avis divergent. Personnellement, je pars du principe qu'il vaut mieux conserver une valeur fausse qu'éliminer une valeur juste. Pour mémoire, la "correction" des données pour les faire cadrer avec les théories du moment peuvent faire passer les scientifiques à côté d'une nouvelle théorie. L'exemple le plus flagrant est la théorie de la relativité : en mécanique classique, 150 000 km/s +150 000 km/s=300 000 km/s. Dans la réalité, les physiciens trouvaient 150 000+150 000=290 000. Beaucoup "arrondissaient" 290 000 à 300 000, considérant 290 000 comme une sorte de valeur aberrante parce qu'elle ne cadrait pas avec la théorie classique.

[7] Vous ne savez pas ce que c'est ? Moi non plus. Et là est bien tout le problème du statisticien...

Cet innocent "arrondi" les faisait passer à côté de la découverte de la relativité générale ! Donc dans le doute, il vaut mieux conserver les données inchangées.

– Enfin, lorsqu'il a un doute ou des raisons de douter, le statisticien peut choisir d'analyser ses données avec des outils "résistants" aux aberrations. De tels outils existent, ils ne nécessitent pas spécialement de connaître la ou les données suspectes, ils sont moins puissants que les outils standards. Ils sont donc à réserver pour les cas où il y a un doute.

Pour mémoire, même si ça n'est pas applicable dans une analyse univarié, la meilleure gestion des valeurs aberrantes consiste a faire deux fois l'analyse statique : une fois avec les données intégrales, une seconde fois après élimination des valeurs aberrantes. Des résultats identiques dans les deux cas donnent une grande crédibilité à l'étude.

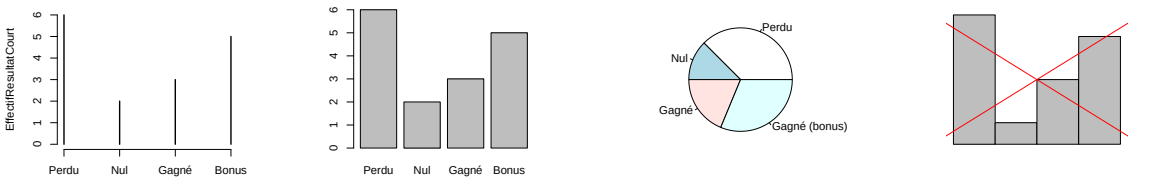
3.2.8 Données extrêmes

Pour finir avec les valeurs aberrantes, il est très important de bien les distinguer des valeurs extrêmes : une valeur aberrante est une valeur liée à une erreur de mesure ou à un hasard extraordinaire qui risque de fausser toute l'expérience (on mesure les performances de gens "normaux" et par hasard, il y a dans la salle un athlète olympique...[[[Histoire des épinard]]]) En particulier, une valeur aberrante N'EST PAS une valeur qui dérange un peu notre hypothèse... Ç à l'air évident, mais en pratique, la tentation peut être forte, lors du nettoyage des données (exclusion des valeurs aberrantes) d'enlever aussi les points un peu excentrés, sorte de "valeurs aberrantes de circonstance"... Bref, chez le statisticien, l'utilisation du proverbe "Quand on veut noyer son chien, on l'accuse de la rage" (ou encore "Quand on veut supprimer des valeurs extrêmes, on les accuse d'être aberrantes") est une faute.

3.2.9 Représentation graphique

On représente les distributions des variables qualitatives par des histogrammes (des graphiques en bâton) en mettant les observations sur l'axe des x ou par des camemberts :

Il existe plusieurs types d'histogramme. Pour les variables qualitatives et discrètes, il est important que les colonnes soient disjointes pour justement montrer qu'entre deux catégories, rien n'existe (un match peut être perdu ou nul, mais pas a mi chemin entre perdu et nul...).

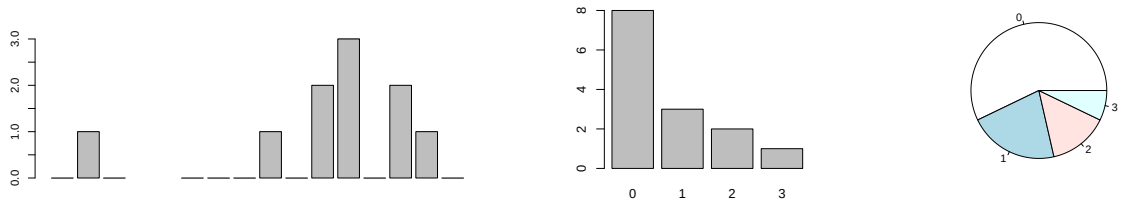


Histo barre Histogramme Camembert Mauvais histogramme

FIG. 3.2.9.a – RCV, différentes représentations graphiques de [RESULTAT]

Pour les variables discrètes, on peut parfois être amené à représenter la liste brute (comme nous l'avons fait pour les notes du contrôle continu). Chaque colonne représente un individu, la hauteur de la colonne représente la valeur de la variable. Ici, on constate que le nombre de cartons jaunes augmente au fur et à mesure que la saison avance.

On peut également représenter l'histogramme, comme pour les variables qualitatives (figure 3.2.9.b)



Liste Brute Histogramme Camembert

FIG. 3.2.9.b – RCV, différentes représentations graphiques de [CARTON]

3.2.10 Représentation graphique des variables continues

La représentation graphique des variables continues est un peu à part de celle des autres variables. En effet, leur nature même (nombre infini ou au moins très grand de modalité possible) rend le comptage de chaque modalité peu informatif, vue que chaque modalité n'est présente qu'un faible nombre de fois. Pour la variable [POINT], la modalité 0 a pour effectif 2, toutes les autres modalités ont pour effectif 1... L'histogramme classique tel que nous l'avons défini pour la variable discrète ne nous apporte donc rien :

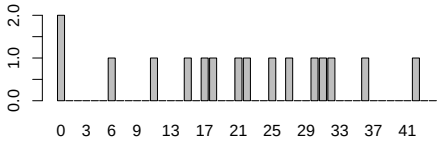
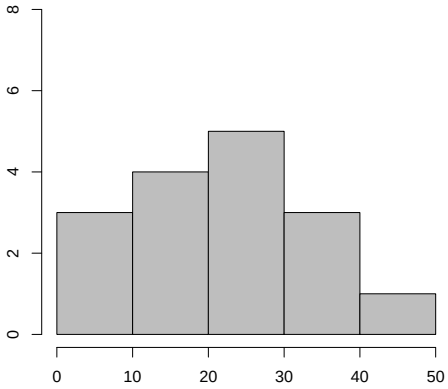


FIG. 3.2.10.a – RCV, histogramme *non regroupé* d'une variable continue

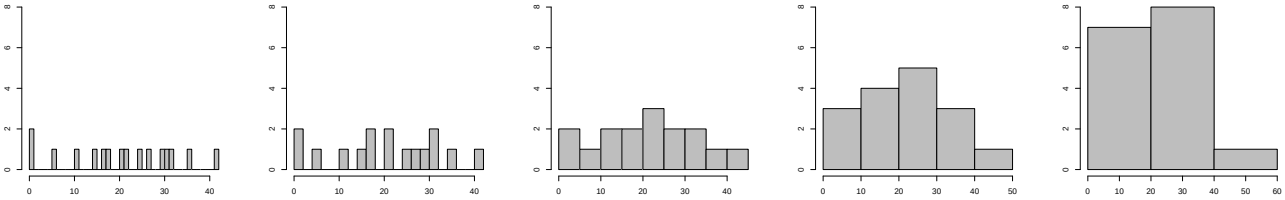
On va donc, au lieu d'utiliser un histogramme avec une colonne par modalité, grouper les modalités dans des intervalles et compter le nombre d'observation contenu dans l'intervalle. Par exemple, on peut choisir de grouper les modalités 10 par 10. Nous compterons donc les observations contenues dans [0 ; 9], dans [10 ; 19], dans [20 ; 29] dans [30 ; 39] et dans [40,49]. Au final, on obtient :

Variable	Intervalles : Effectifs
[Point]	$[0;9]$: 3 $[10;19]$: 4 $[20;29]$: 5 $[30;39]$: 3 $[40;49]$: 1



Tab. 3.2.10.b – Effectifs et histogramme des intervalles

D'autres découpages sont possibles : nous aurions également choisir des intervalles de taille 2, 5 ou même 20. Le choix dépend de ce que l'on cherche à mettre en évidence. Il est clair que plus on regroupe, plus on perd de l'information (l'histogramme groupe de 20 ne nous apprend pas grand chose), mais si on ne regroupe pas assez, trop d'informations empêchent d'avoir un regard synthétique sur les données.



Non regroupé | Groupe de 2 | Groupe de 5 | Groupe de 10 | Groupe de 20

Fig. 3.2.10.c – Différents regroupements possibles pour [Points]

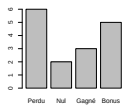
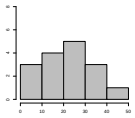
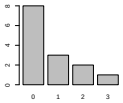
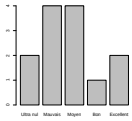
En pratique, les logiciels statistiques choisissent automatiquement le nombre d'intervalle de la variable (naturellement, on peut modifier leur choix s'il ne nous plait pas...)

Attention Piège : Comme nous l'avons vu dans le paragraphe 2.2.4, il est possible de transformer une variable continue en variable discrète ou en une variable ordonnée, exactement selon le même principe que celui que nous venons d'utiliser. Cependant, dans le cas présent, cette transformation ne sert qu'à la représentation graphique. En particulier, dans la suite de l'étude, [Point] reste une variable continue et lorsque nous aurons à calculer des indices, nous utiliserons les vraies observations de point. Il est donc important de distinguer le recodage d'une variable continue en une variable discrète ou ordonnée (la nature de la variable change) et une transformation momentanée ayant pour seul but une représentation graphique (la nature de la variable ne change pas.)

[[[Distribution théorique : vrai pour les discrètes, compliqué pour les continues]]]

3.2.11 Récapitulatif

Nous pouvons maintenant présenter un tableau complet résumant les variables et leurs effectifs :

Variable	Nature	Effectifs	Manquantes
[Resultat]	Ordonnée	Perdu : 6 (37,5%) Nul : 2 (12,5%) Gagné : 3 (18,75%) Gagné (bonus) : 5 (31,25%) 	0
[Points]	Continue	$[0; +\infty[$ 	0
[Carton]	Discrète	0 ; 1 ; 2 ; ... 	0
[Meteo]	Nominale	Soleil : 9(56,25%) Pluie : 3(18,75%) Nuageux : 3(18,75%) Neige : 1(6,25%) 	0
[Commentaire]	Ordonnée	Ultra nul : 2(20%) Mauvais : 4(20%) Moyen : 4(20%) Bon : 1(20%) Excellent : 2(20%) 	3

TAB. 3.2.11.a – RCV, récapitulatif

3.3 Indice de centralité

Les indices de centralité permettent de définir approximativement le "centre" des observations. Différentes définitions de centre ont donnée naissance à différents indices parmi lesquels on trouve le mode, la médiane et la moyenne.

3.3.1 Le mode

Le mode est l'indice de centralité le plus facilement calculable. C'est simplement l'observation ayant le plus grand effectif.

Définition 3.3.1.a : Mode
Le **Mode** (ou **Valeur dominante**) est l'observation la plus fréquente.

Chez Harris, le mode de la variable [Reponse] est (Non). Concernant le Racing Club Villeneuvois, le mode de [Resultat] est (Perdu). Le mode d'une distribution n'est pas forcément unique. Si une distribution n'a qu'un mode, elle est unimodale. Si elle en a plusieurs, elle est bimodale, trimodale et ainsi de suite.

En pratique, le mode est utilisé principalement sur des variables nominales (pas vraiment par choix, mais plus parce que sur ce type de variable, c'est le seul indice de centralité disponible...) Même dans ce cadre, il est a manipuler avec précaution tant sa volatilité est grande. L'industrie agro-alimentaire présente un parfait exemple de cette faiblesse. La loi oblige les producteurs de denrées alimentaires à inscrire sur leurs marchandises la liste des ingrédients, et cela par ordre d'importance décroissante. Traduit en vocabulaire statistique, cela signifie que le mode doit être en tête de liste. Or, nombre de produit fortement sucrés, comme par exemple les confitures, tentent de prétendre qu'ils sont riches en fruit et faiblement sucré. D'où l'idée de ne pas considérer le sucre comme un ingrédient mais de le séparer en deux, par exemple glucose et fructose. Le mode passe donc de sucre à fruit :

Composant	Pourcentage
Sucre (mode)	50%
Fruit	35%

$\Longrightarrow$

Composant	Pourcentage
Fruit (mode)	35%
Glucose	27%
Fructose	23%

TAB. 3.3.1.a – Comment changer le Mode des denrées alimentaires

Par un tour de passe-passe, le mode vient de sauter de sucre à fruit... Ingénieux, n'est ce pas ?

3.3.2 La médiane

La médiane est un indice calculable sur une variable disposant d'une relation d'ordre (donc qualitative ordonnée ou quantitative). Intuitivement, c'est assez simple : la médiane est l'observation qui divise la population en deux groupes : 50% des individus ont une valeur inférieure à la médiane, 50% ont une valeur supérieure. Pour la trouver, il suffit de ranger les individus par ordre croissant, la médiane est alors l'observation prise par l'individu du milieu, soit celui de rang $\frac{n+1}{2}$.

3.3.2.1 Variable qualitative

Cette définition est un peu théorique, mais le concept en lui-même est très simple : considérons la variable [COMMENTAIRE]. C'est une variable ordonnée : (Ultra nul) < (Mauvais) < (Moyen) < (Bon) < (Excellent). Classons les observations de cette variable par ordre croissant :

1	2	3	4	5	6	7	8	9	10	11	12	13
Ultra nul	Ultra nul	Mauvais	Mauvais	Mauvais	Mauvais	*Moyen*	Moyen	Moyen	Moyen	Bon	Excellent	Excellent

Il y a 13 individus, la médiane est donc l'observation de rang $\frac{13+1}{2} = 7$ soit l'observation (*Moyen*).

Malheureusement, cette définition ne peut s'appliquer que lorsque n est impair. En effet, pour la variable [RESULTAT], il y a 16 observations :

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Perdu	Perdu	Perdu	Perdu	Perdu	Perdu	Nul	Nul	*Gagné*	Gagné	Gagné	Bonus	Bonus	Bonus	Bonus	Bonus

Idéalement, on aimerait pouvoir prendre pour médiane la valeur correspondant à $\frac{16+1}{2} = 8,5$. Mais cette observation n'existe pas. On doit donc choisir une observation voisine de 8,5. Par convention, on prend l'observation immédiatement supérieure. La médiane est donc l'observation de rang $\frac{16+2}{2} = 9$ soit l'observation (∗Gagne∗).

Définition 3.3.2.a : Médiane (qualitative)
Soit une population de n individus et une variable qualitative ordonnée. Considérons les individus classés par ordre croissant. La **Médiane** est l'observation prise par l'individu de rang $\frac{n+1}{2}$ si n est impair, $\frac{n+2}{2}$ si n est pair.

3.3.2.2 Variable quantitative

Pour les variables quantitatives, la définition de la médiane est la même dans le cas d'une variable impaire : l'observation prise par l'individu de rang $\frac{n+1}{2}$. Par contre, dans le cas d'une variable paire, les propriétés arithmétiques des variables quantitatives permettent de ne pas avoir à choisir entre deux observations mais de prendre la valeur qui se situe *entre* les observations. Pour la variable [POINTS], on a :

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	0	6	11	15	17	18	21	22	25	27	30	31	32	36	42

La médiane se situe donc entre les observations des individus de rang 8 et 9, soit entre (21) et (22). Que prendre : 21 ? 22 ? 21 ? 22 ? Les statisticiens ont tranché, ils lui attribuent la valeur située à mi chemin entre 21 et 22, soit 21,5, la moyenne[8]de 21 et 22 :

Définition 3.3.2.b : Médiane (quantitative)
Soit une population de n individus et une variable quantitative. Considérons les individus classés par ordre croissant. La **Médiane** est l'observation prise par l'individu de rang $\frac{n+1}{2}$ si n est impair, la moyenne des observations de rang $\frac{n}{2}$ et $\frac{n+2}{2}$ si n est pair.

Au final, la variable [POINTS] a pour médiane $\frac{21+22}{2} = 21,5$.

3.3.3 Moyenne

La moyenne (de son vrai nom : **Moyenne arithmétique** est l'indice de centralité le plus usité, tout le monde l'a déjà manipulé simplement parce que c'est l'indice utilisé pour déterminer si oui ou non un élève accède à la classe supérieure. Une moyenne générale égale ou supérieur à 10 est synonyme de succès. On peut l'interpréter comme la valeur la plus proche de toutes les observations.

Définition 3.3.3.a : moyenne
La **Moyenne** (ou **Moyenne arithmétique**) d'une variable quantitative [X] est la somme des observations divisé par le nombre total d'observation. On la note $\overline{X}$.
$\overline{X} = \frac{\sum x_i}{n}$

Le calcul de la moyenne nécessite des additions. Cet indice n'est donc évaluable que pour des variables quantitatives.

- La moyenne de 3,4 et 5 est $\frac{3+4+5}{3} = 4$
- La moyenne de [CARTONS] est $\frac{0+1+0+0+1+0+0+0+1+0+2+3+0+2+1+0}{16} = \frac{11}{16} = 0,69$
- $[\overline{Points}] = \frac{0+32+25+11+30+42+6+22+21+31+17+0+15+27+18+36}{16} = 20,81$

3.3.4 Récapitulatif

Applique au Racing Club Villeneuvois, cela donne

Variable	Nature	Mode	Médiane	Moyenne
[RESULTAT]	Ordonnée	Perdu	Gagné	<ND>
[POINTS]	Continue	<SI>	22,5	20,81
[CARTONS]	Discrète	0	0	0,69
[METEO]	Nominale	Soleil	<ND>	<ND>
[COMMENTAIRE]	Ordonnée	Mauvais, Moyen	Moyen	<ND>

TAB. 3.3.4.a – RCV, mode / médiane / moyenne

Comme l'indique le tableau, pour la majorité des variables, plusieurs indices sont disponibles. Se pose alors la question du choix du meilleur indice...

3.3.5 Le Best Centrality Award...

Nous disposons donc de trois indices de centralité. La moyenne est clairement l'indice le plus utilisé. Est-ce justifié ? Il ne nous appartient pas de trancher. Par contre, nous pouvons vérifier les propriétés de ces indices, les comparer, vérifier leurs avantages et inconvénients. Une sorte de compétition inter-indice...

[8] La définition formelle de la médiane nécessite de faire appel à la moyenne. En toute rigueur, nous devrions donc définir la moyenne avant la médiane. Cependant, le concept intuitif de médiane étant plus simple que celui de moyenne et la moyenne de deux nombres étant tout de même un concept connu de la majorité, nous privilégions ici l'ordre pédagogique à l'ordre formel...

Les nominés sont :

- Le mode
- La médiane
- La moyenne

Ils concourent dans les catégories :

- Cohérence définitoire
- Suffisance
- Absence de biais
- Efficacité
- Robustesse
- Polyvalence
- Simplicité d'évaluation

3.3.5.1 Cohérence définitoire

- Le **mode** se concentre sur la valeur la plus fréquente. Dans le cas de la variable `[Point]`, cette valeur est zéro. Pourtant, si l'on devait résumer l'ensemble de la saison pas un seul nombre, zéro n'est certes pas celui qui viendrait à l'esprit. Le mode ne respecte pas la cohérence définitoire ✗

- Comme nous l'avons déjà signalé, la cohérence, concept fondamental, est néanmoins subjective. La **médiane** et la **moyenne** sont toutes deux considérées comme représentant raisonnablement le concept de centre[9]. Elles sont cohérentes. ✓

3.3.5.2 Suffisance

- Le **mode** se concentre sur les valeurs les plus fréquentes, il ignore toutes les autres. Le mode de `[Resultat]` = { `(Perdu,5)` ; `(Nul,2)` ; `(Gagné,3)` ; `(Gagné (bonus),4))` } est `(Perdu)`. Si un `(Nul)` devient `(Gagné)`, le mode ne change pas du tout. Par contre, si un simple `(Perdu)` de moins et un `(Gagné (bonus))` et le mode passerait d'un extrême à l'autre, de `(Perdu)` à `(Gagné (bonus))` (alors qu'un unique match aurait une importance mineure sur la saison) : le mode n'est pas suffisant ✗

- La **médiane** classe les observations pour ne conserver que celles du centre. Tout comme le mode, elle ignore la grande partie des valeurs : la médiane n'est pas suffisante ✗

- La **moyenne** additionne TOUTES les observations avant de les diviser par l'effectif : la moyenne est suffisante ✓

3.3.5.3 Absence de biais

Le mode, la médiane et la moyenne sont tous des indices non biaisés ✓. La démonstration de cette propriété est un peu délicate, nous ne la donnerons pas ici. Toutefois, nous présenterons sur un exemple une vérification qui donnera une bonne idée de la démarche générale en annexe, section **??** Moyenne non biaisée.

3.3.5.4 Efficacité

Comme l'absence de biais, l'efficacité est difficile à démontrer. Intuitivement, elle est lié à la sensibilité aux variations. Considérons une population dans son ensemble. Si un indice est stable, le fait d'enlever des individus ne changera que peu sa valeur. S'il est instable, un individu ôté en trop et l'indice bascule.

- Le mode est très instable donc peu efficace ✗.
- La moyenne est très efficace ✓.
- La médiane est efficace (un peu moins que la moyenne, mais cela reste raisonnable) ✓.

Concrètement, cela signifie que pour évaluer la moyenne d'une population, un petit échantillon donnera une bonne précision. Pour le mode, il faudra nécessairement un échantillon important.

3.3.5.5 Robustesse

La robustesse est directement opposée à la suffisance. On ne s'étonnera donc pas de constater que le mode et la médiane sont robustes, la moyenne ne l'est pas. [[[Exemple (référence) à fouiller]]] La biologie

[9] Vous n'êtes pas d'accord ? C'est votre droit. Refusez, tempêtez, vitupérez, remettez en cause. Peut être de votre réflexion naîtra un nouveau concept qui relèguera la moyenne au rang des indices périmés.

animale permet de bien illustrer le concept : le meilleur ami du chercheur, comme chacun sait, est le rat. Dans de nombreuses expériences, des chercheurs mesurent le temps que des rats mettent à traverser un labyrinthe, pour exemple pour étudier l'effet d'apprentissage. Lors d'un premier passage, ils ont mis un temps moyen de 22,6 secondes et un temps médian de 20 secondes. Lors du deuxième passage, les dix rats obtiennent les performances suivantes (mesurées en seconde) : 17, 16, 19, 24, 16, 13, 15, 16, 5949 et 19. Le temps moyen pour sortir du labyrinthe est 610,4 secondes, soit 10 minutes et 10 secondes. Le temps médian est 16,5 secondes... Si on considère la médiane, il y a un effet apprentissage puisque le temps médian est passé de 20 à 16,5 secondes. Selon la moyenne, le deuxième passage est très lent. Après examen des données, cette forte moyenne est lié à la performance du neuvième rat : il met 5949 secondes à sortir. Comment une telle valeur est-elle possible ? Erreur de notation ? "Non, nous a répondu l'auteur, c'est simplement un rat qui s'est endormi dans le labyrinthe"[10] !

Retour à nos indices. Un rat s'endort :

- La moyenne est complètement perturbée, elle n'est pas robuste ✗.
- La médiane ne bouge pas, elle est robuste ✓.
- Le mode n'aurait pas bougé non plus ✓.

3.3.5.6 Polyvalence

- Le mode ne nécessite pas d'addition ou de comparaison. Il peut donc se calculer sur tous les types de variables. Par contre, sur des variables continues, il est extrêmement sensible au découpage de la distribution : le mode de `[POINT]` est zéro, si on regroupe les valeurs par 5 ou 10, le mode change complètement. De fait, le mode ne présente aucun intérêt pour les variables continues. Il est raisonnable polyvalent (3 sur 4). ✓
- La médiane se calcule à l'aide de comparaison mais sans addition. On peut donc l'évaluer sur les ordonnées et les quantitatives. Elle est raisonnablement polyvalente (3 sur 4). ✓
- La moyenne nécessite des additions. On ne peut l'utiliser que sur des variables quantitatives. Elle est modérément polyvalente (2 sur 4). ✗

En pratique, l'importance de ce critère varie selon les domaines. En médecine, en biologie et autre sciences de la vie, l'usage des variables nominales est plutôt rare. La polyvalence de la médiane est donc suffisante.

3.3.5.7 Simplicité d'évaluation

Comme nous l'avons déjà précisé section??, cette rubrique n'a plus aucune raison d'être : toutes les analyses statistiques se font avec ordinateur. Qu'il mette 0,053 ou 0,535 secondes à faire les calculs, cela n'a strictement aucune importance... Pour information, le mode était facile à calculer, la médiane un peu moins et la moyenne encore moins (sur des petits jeux de données, la différence n'est pas évidente. Mais si un jour vous devez calculez à la main la moyenne ou le mode de 10 000 prix, choisissez le mode...)

3.3.5.8 Bilan

[[[A ajouter : avantage de la médiane, sa valeur est réalisée. Cas où on pourrait l'utiliser : pour les concours ; au lieu de dire "faut une moyenne de tant", moyenne qui change en fonction des autres, on pourrait dire..." heu... A fouiller ! Autre truc à fouiller : la France est au 4ème rang des consommateurs de truc muche en Europe]]]
En pratique,

- Le mode n'a pratiquement que des inconvénients...
- La médiane a pour faiblesse sa non suffisance. Dans l'esprit des gens, il peut sembler "injuste" que le résumé des valeurs ne prennent pas en compte toutes les valeurs (la médiane de 3, 8, 9 est la même que celle de 5, 8, 15 ou de 8, 8, 15... Si vous faites redoubler 8, 8 , 15 au même titre que 3, 8, 9, préparez vous à recevoir les parents.
- Comme nous l'avons vu avec les rats qui s'endorment, la moyenne a pour faiblesse sa grande sensibilité aux valeurs aberrantes ou extrêmes. Cette faiblesse peut parfois se corriger en excluant les valeurs aberrantes (comme décrit section 3.2.7)

[10] Si cet exemple est authentique, nous ne pouvons néanmoins pas fournir une référence précise d'article, vu que ce genre de données sont effacées avant publication... Mais selon les biologistes, cela arrive régulièrement

Le mode est donc utilisé seulement pour les variables nominales.
La médiane est calculable sur des variables ordonnées. Dans ce cadre, c'est donc elle qui sera choisie.
Concernant les variables continues, entre médiane et moyenne, le choix semble difficile. En fait, c'est l'efficacité qui va trancher. Dans les essais médicaux, inclure un patient coûte cher (cela va jusqu'à 10000€par patient). Dans l'enseignement, corriger un paquet de copie coûte TRÈS cher (en temps). En psychologie, faire se déplacer les gens pour participer à une expérience est également assez difficile. Le nombre de sujet est donc un critère primordial. La moyenne étant plus efficace que la médiane, c'est elle qui sera choisie. Mais ça n'empêche pas certains auteurs de travailler avec la médiane, en particulier quand le risque de valeur aberrante est important.
La table 3.3.5.a résume la compatibilité entre les variables et les indices de centralité (avec <ND> = Non Défini, <SI> = Sans intérêt, le nombre d'étoiles indiquant l'intérêt de l'indice) :

Variable	Mode	Médiane	Moyenne
Nominale	Oui*	<ND>	<ND>
Ordonnée	Oui	Oui***	<ND>
Discrète	Oui	Oui**	Oui***
Continue	<SI>	Oui**	Oui***

TAB. 3.3.5.a – Compatibilité entre les variables et les indices de centralité

3.4 Indices de dispersion

Les indices de dispersion donne une idée du regroupement ou de l'éparpillement des données autour de l'indice de centralité.

3.4.1 Dispersion pour les variables nominale

Nous avons ajouté ce paragraphe uniquement pour vous éviter de vous demander s'il n'avait pas été oublié, comme il est d'ailleurs oublié dans tous les livres de statistique. Ca n'est pas un oubli : il n'existe pas d'indice de dispersion pour les variables nominales... Problème réglé!

3.4.2 Minimum et maximum

Le minimum et le maximum sont calculables sur les variables ordonnées ou quantitatives.

Définition 3.4.2.a : Minimum
Le **Minimum** d'une variable ordonné ou quantitative est l'observation la plus petite.

Définition 3.4.2.b : Maximum
Le **Maximum** d'une variable ordonné ou quantitative est l'observation la plus grande.

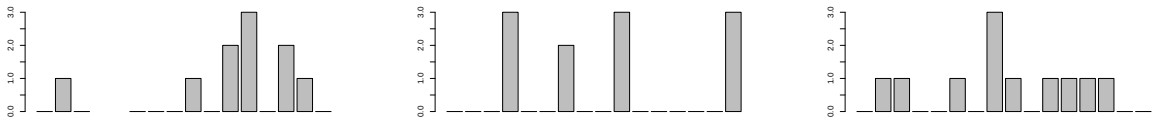
Appliqué au Racing Club Villeneuvois, on obtient :

Variable	Nature	Minimum	Maximum
[RESULTAT]	Ordonnée	Perdu	Gagné (bonus)
[POINTS]	Continue	0	42
[CARTON]	Discrète	0	3
[METEO]	Nominale	<ND>	<ND>
[COMMENTAIRE]	Ordonnée	Ultra nul	Excellent

TAB. 3.4.2.a – RCV, minimum / maximum

Ces deux indices donnent une idée relativement grossière de la dispersion. En particulier, sur une saison, il est probable que tous les clubs du Comminges auront perdu une fois, gagné une fois avec bonus, ils auront donc tous les mêmes minimum et maximum. De même, le Stade de Labarthe de Rivière a une variable [CARTONS] qui vaut 0, 0, 0, 3, 0, 0, 2, 0, 0, 3, 0, 0, 0, 0, 0, 3 alors que pour La serre [CARTON]={0,

1, 1, 0, 0, 1, 0, 3, 1, 0, 1, 1, 1, 1, 0, 0}. Les moyennes, minimum et maximum des trois équipes sont identiques. Pourtant, quand on représente graphiquement les données brutes, les trois variables cartons ne se ressemblent pas.



Villeneuve Labarthes La Serre

FIG. 3.4.2.b – Cartons, comparaison des listes brutes

D'où l'introduction de *points de contrôle* intermédiaires, les quartiles.

3.4.3 Les quartiles

Les quartiles et les quantiles sont, comme la médiane, des indices calculables sur une variable disposant d'une relation d'ordre (donc qualitative ordonnée ou quantitative). Intuitivement, c'est assez simple :

- Pour mémoire, la médiane était l'observation qui divisait population en deux groupes : 50% des individus ont une valeur inférieure à la médiane, 50% ont une valeur supérieure.
- De manière similaire, le premier quartile est l'observation qui divise la population en deux groupes : 25% des individus ont une valeur inférieure au premier quartile, 75% ont une valeur supérieure.
- Enfin, le troisième quartile est l'observation qui divise la population en deux groupes : 75% des individus ont une valeur inférieure au troisième quartile, 25% ont une valeur supérieure.

Pour les trouver, il suffit de ranger les individus par ordre croissant, le premier quartile est l'observation prise par l'individu de rang $\frac{n+3}{4}$, le troisième est l'observation prise par l'individu de rang $\frac{3n+1}{4}$ (le détail de ses formules est donnée section 3.4.3.2). Comme pour la médiane, il va nous falloir distinguer les variables ordonnées des quantitatives.

3.4.3.1 Variable qualitative ordonnée

Pour la variable [COMMENTAIRES], cela correspond aux observations de rang $\frac{13+3}{4} = 4$ et $\frac{3\times13+1}{4} = 10$ soit :

1	2	3	4	5	6	7	8	9	10	11	12	13
Ultra nul	Ultra nul	Mauvais	*Mauvais*	Mauvais	Mauvais	Moyen	Moyen	Moyen	*Moyen*	Bon	Excellent	Excellent

Naturellement, les problèmes qui se sont posés avec la médiane vont se poser ici aussi, en particulier le problème des arrondis lorsque de calcul ne tombe pas juste. Plusieurs cas peuvent se présenter :

- Si le premier quantile a une position qui se termine en 0,25 on choisi l'observation inférieure.
- Si le premier quantile a une position qui se termine en 0,5 on choisi par convention [[[à vérifier]]]l'observation inférieure.
- Si le premier quantile a une position qui se termine en 0,75 on choisi l'observation supérieur.
- Si le troisième quantile a une position qui se termine en 0,25 on choisi l'observation inférieure.
- Si le troisième quantile a une position qui se termine en 0,5 on choisi par convention [[[à vérifier]]]l'observation supérieure.
- Si le troisième quantile a une position qui se termine en 0,75 on choisi l'observation supérieur.

Pour la variable [RESULTAT], il y a 16 observations. Les premiers et troisièmes quartiles sont donc situés au rang $\frac{16+3}{4} = 4,75$ que l'on arrondi à 5, et a u rang $\frac{3\times16+1}{4} = 12,25$ que l'on arrondi à 12 :

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Perdu	Perdu	Perdu	Perdu	*Perdu*	Perdu	Nul	Nul	Gagné	Gagné	Gagné	*Bonus*	Bonus	Bonus	Bonus	Bonus

Formaliser cette définition nécessite l'introduction de deux fonctions : $Arrondi\_Inf$ est la fonction partie entière inférieure. Elle arrondit un nombre à l'entier le plus proche, et quand il y a hésitation (comme

2,5, on hésite entre 2 et 3), elle choisi la plus petite valeur (ici : 2). Donc $2 \longrightarrow 2$; $2,25 \longrightarrow 2$; $2,5 \longrightarrow 2$; $2,75 \longrightarrow 3$; $3 \longrightarrow 3$.
A l'opposé, *Arrondi_Sup* arrondit un nombre à l'entier le plus proche, et quand il y a hésitation, elle choisi la plus grande valeur. Donc $2 \longrightarrow 2$; $2,25 \longrightarrow 2$; $2,5 \longrightarrow 3$; $2,75 \longrightarrow 3$; $3 \longrightarrow 3$.

Définition 3.4.3.a : Premier quartile (qualitatif)
Soit une population de n individus et une variable qualitative ordonnée. Considérons les individus classés par ordre croissant. Le **Premier quartile (qualitatif)** est l'observation prise par l'individu de rang $Arrondi\_Inf\left(\frac{n+3}{4}\right)$.

Définition 3.4.3.b : troisième quartile (qualitatif)
Soit une population de n individus et une variable qualitative ordonnée. Considérons les individus classés par ordre croissant. Le *troisième quartile* est l'observation prise par l'individu de rang $Arrondi\_Sup\left(\frac{3n+1}{4}\right)$.

Au final, la notion de quartile se généralise, le minimum, le maximum et la médiane n'étant que des quartiles particuliers :

- $Q0$: le quartile zéro (ou minimum) est l'observation qui isole les 0% inférieurs des individus.
- $Q1$: le premier quartile est l'observation qui isole les 25% inférieurs des individus.
- $Q2$: le deuxième quartile (ou médiane) est l'observation qui isole les 50% inférieurs des individus.
- $Q3$: le troisième quartile est l'observation qui isole les 75% inférieurs des individus.
- $Q4$: le quatrième quartile (ou maximum) est l'observation qui isole les 100% inférieurs des individus.

3.4.3.2 Variables quantitatives

Dans le cas de variables quantitatives, le concept est le même (séparation de la population en tranche de 25%) mais le calcul est un peu plus délicat. Considérons $RangQ1$, le rang du premier quartile :

- $RangQ1$ est à mi-chemin entre l'individu de rang 1 et l'individu médian,
- $RangQ1$ est donc à mi-chemin entre l'individu de rang 1 et l'individu de rang $\frac{n+1}{2}$.
- $RangQ1$ est donc la moyenne du rang 1 et $\frac{n+1}{2}$ c'est-à-dire $RangQ1 = \frac{1+\frac{n+1}{2}}{2} = \frac{n+3}{4}$.

De la même manière, $RangQ3$ est à mi-chemin entre la médiane et l'individu de rang n, soit $RangQ3 = \frac{\frac{n+1}{2}+n}{2} = \frac{3n+1}{4}$.

Définition 3.4.3.c : Rang des quartiles
Soit une population de taille n.

- **RangQ0**, le rang du quartile zéro (ou minimum), est 1
- **RangQ1**, le rang du premier quartile, est $\frac{n+3}{4}$
- **RangQ2**, le rang du deuxième quartile (ou médiane), est $\frac{n+1}{2}$
- **RangQ3**, le rang du troisième quartile, est $\frac{3n+1}{4}$
- **RangQ4**, le rang du quatrième quartile (ou maximum), est n

Retour à notre premier quartile. Quand $RangQ1$ est une valeur entière, le quartile est simplement l'observation correspondante. Quand $RangQ1$ n'est pas entier, le quartile est évalué par **interpolation linéaire**.

Prenons 5 exemples, considérons les variables suivantes :

- $[A_5]$ sur 5 individus : {6, 6, 8, 8, 8} donc $RangQ1$ vaut $\frac{5+3}{4} = 2$
- $[A_6]$ sur 6 individus : {6, 6, 8, 8, 8, 8} donc $RangQ1$ vaut $\frac{6+3}{4} = 2,25$
- $[A_7]$ sur 7 individus : {6, 6, 8, 8, 8, 8, 8} donc $RangQ1$ vaut $\frac{7+3}{4} = 2,5$
- $[A_8]$ sur 8 individus : {6, 6, 8, 8, 8, 8, 8, 8} donc $RangQ1$ vaut $\frac{8+3}{4} = 2,75$
- $[A_9]$ sur 9 individus : {6, 6, 8, 8, 8, 8, 8, 8, 8} donc $RangQ1$ vaut $\frac{9+3}{4} = 3$

Pour chacune d'entre elles, évaluons Q1 :

- Pour $[A_5]$, Q1 est simplement la valeur de l'observation 2, soit `(6)`.
- Pour $[A_9]$, Q1 est simplement la valeur de l'observation 2, soit `(8)`.
- Pour les autres, la valeur de Q1 est choisi entre 6 et 8 de manière proportionnelle :
  - Pour $[A_6]$ (position 2,25), Q1 sera plus proche de 6 que de 8
  - Pour $[A_7]$ (position 2,5), Q1 sera exactement entre 6 et 8

– Pour [A6] (position 2,75) Q1 sera proche de 8

Cette manière de choisir est l'interpolation linéaire.

L'interpolation linéaire consiste, lorsqu'on hésite entre deux valeurs (par exemple 6 et 8 dans nos exemples), à tracer une droite entre les valeurs et à choisir comme quartile l'image de son rang par la droite.

– Pour [A6] (position 2,25), Q1 sera l'image de 2,25 par la droite, soit 6,5
– Pour [A7] (position 2,5), Q1 sera l'image de 2,5 par la droite, soit 7
– Pour [A8] (position 2,75), Q1 sera l'image de 2,75 par la droite, soit 7,5

La figure 3.4.3.c résume la manière de calculer Q1 :

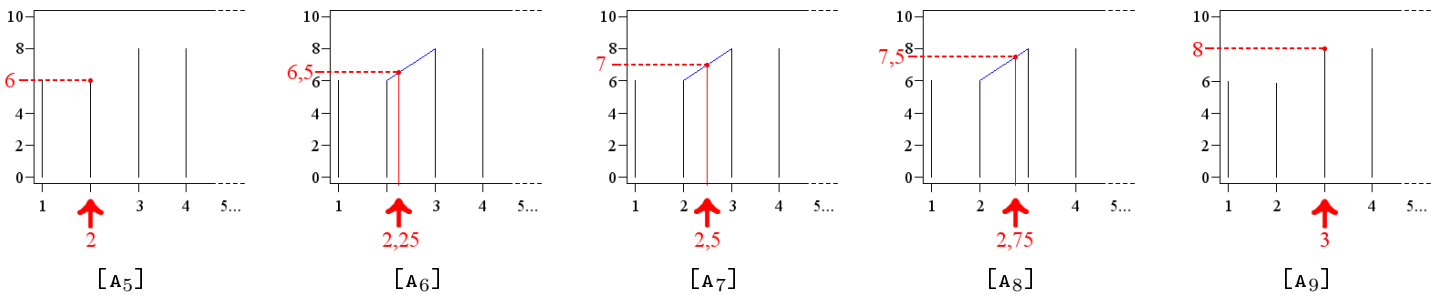


FIG. 3.4.3.c – Interpolation linéaire

Pour [A6], le $RangQ1$ est 2,25, le premier quartile est donc 6,5. Le quartile est donc une sorte de moyenne (moyenne pondérée pour être précis) entre ses deux valeurs voisines, pondérations qui vont dépendre des décimales (0,25 ; 0,5 ou 0,75) du rang du cartile. On définit donc la fonction Dec, fonction qui donne les décimales d'un nombre : $Dec(2) = 0$; $Dec(2,25) = 0,25$; $Dec(2,5) = 0,5$ et $Dec(2,75) = 0,75$

Nous pouvons maintenant définir formellement le premier quartile :

Définition 3.4.3.d : Premier quartile (quantitatif)
Soit une population de n individus et une variable quantitative. Considérons les individus classés par ordre croissant.

– Si $RangQ1$ est une valeur entière, le **Premier quartile** est l'observation prise par l'individu de rang $RangQ1$.
– Si $RangQ1$ n'est pas une valeur entière :
  – Soit $Q1^-$ l'observation dont le rang est immédiatement inférieure ou égal à $RangQ1$ et $Q1^+$ l'observation dont le rang est immédiatement supérieure ou égal à $RangQ1$.
  – Alors le premier quartile a pour valeur la moyenne de $Q1^-$ et $Q1^+$ pondérés respectivement par les coefficients $(1 - Dec(RangQ1)$ et $Dec(RangQ1)$.

$$Q1 = (1 - Dec(RangQ1)) \times Q1^- + Dec(RangQ1) \times Q1^+$$

Pour la variable [A5], le $RangQ1 = \frac{5+3}{2} = 2$ est une valeur entière, le quartile est l'observation de rang 2, c'est-à-dire 6.

Pour la variable [A6], le $RangQ1 = \frac{6+3}{2} = 2,25$ n'est pas une valeur entière. $Dec(RangQ1)$ vaut 0,25, l'observation inférieure $Q1^-$ est 6, l'observation supérieure $Q1^+$ est 8. La valeur du quartile est donc $(1 - 0,25) \times 6 + 0,25 \times 8 = 6,5$. Ca marche !

Tout ce que nous venons de voir est rigoureusement symétrique pour le troisième quartile :

Définition 3.4.3.e : Troisième quartile (quantitatif)
Soit une population de n individus et une variable quantitative. Considérons les individus classés par ordre croissant.

- Si $RangQ3$ est une valeur entière, le **Troisième quartile** est l'observation prise par l'individu de rang $RangQ3$.
- Si $RangQ3$ n'est pas une valeur entière :
  - Soit $Dec(Rang3)$ les décimales du rang premier quartile
  - Soit $Q3^-$ l'observation dont le rang est immédiatement inférieure ou égal à $RangQ3$ et $Q3^+$ l'observation dont le rang est immédiatement supérieure ou égal à $RangQ3$.
  - Alors le troisième quartile a pour valeur la moyenne de $Q3^-$ et $Q3^+$ pondérés respectivement par les coefficients $(1 - Dec(RangQ3)$ et $Dec(RangQ3)$.

$$Q3 = (1 - Dec(RangQ3)) \times Q3^- + Dec(RangQ3) \times Q3^+$$

3.4.3.3 Exemple

Retour au Racing Club Villeneuvois. Calculons les quartiles de la variable [POINT] :

- Le rang de Q1 est $RangQ1 = \frac{16+3}{4} = 4,75$. Q1 se situe entre 11 et 15. Plus précisément, $Q1 = 0,25 \times 11 + 0,75 \times 15 = 14$
- Le rang de Q3 est $RangQ3 = \frac{3\times16+3}{4} = 12,25$. Q3 se situe entre 30 et 31. Plus précisément, $Q1 = 0,75 \times 30 + 0,25 \times 31 = 30,25$

1	2	3	4 | * | 5	6	7	8	9	10	11	12 | * | 13	14	15	16
0	0	6	11 | * | 15	17	18	21	22	25	27	30 | * | 31	32	36	42

Variable	Nature	Q0	Q1	Q2	Q3	Q4
[RESULTAT]	Ordonnée	Perdu	Perdu	Gagné	Gagné (bonus)	Gagné (bonus)
[POINTS]	Continue	0	12,25	21,5	30,25	42
[CARTON]	Discrète	0	0	1	1	3
[METEO]	Nominale	<ND>	<ND>	<ND>	<ND>	<ND>
[COMMENTAIRE]	Ordonnée	Ultra nul	Mauvais	Moyen	Moyen	Excellent

TAB. 3.4.3.e – RCV, les quartiles

3.4.3.4 Autres quantiles

De la même manière que les cinq quartiles découpent les individus en groupe de 25%, les onze déciles découpent la population en groupes de 10%, les cent-un centiles la découpe en groupe de 1%... Là encore, quand le rang d'un quantile ne tombe pas juste, on le calcule à l'aide d'une interpolation linéraire. Les principes et les définitions sont similaires à celles des quartiles, nous ne détaillerons pas ici.

3.4.4 L'étendu

L'étendu n'est calculable que sur les variables quantitatives. En effet, son calcul nécessite une soustraction.

Définition 3.4.4.a : Étendue
L'**Étendue** d'une variable quantitative est la distance séparant le minimum du maximum.

$$Etendue = Minimum - Maximum = Q4 - Q0$$

Pour [POINTS], l'étendu est 42-0=42. Pour la variable [B_5] prenant les valeur { 5, 7, 9, 10, 11}, l'étendu est 11-5=6.

L'étendue inter quartile est la distance séparant les quartiles Q1 et Q3. Elle est surtout utilisée dans la construction des boîtes à moustaches (représentation graphique des quartiles, comme nous le verrons

section ??) car elle délimite en espace contenant 50% des observations (25% des observations sont en dessous de Q1, 25% sont au-dessus de Q3 donc 50% sont entre Q1 et Q3)

Définition 3.4.4.b : étendue inter quartile
L'**Étendue inter quartile** d'une variable quantitative est la distance séparant le premier quartile du troisième.

$$EtendueIQ = Q3 - Q1$$

Pour [POINTS], l'étendu inter quartile est 30,25-14=16,25. Pour la variable [B_5], l'étendu inter quartile est 10-7=3.

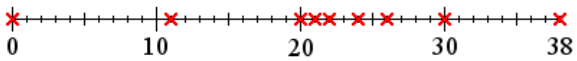
Variable	Nature	Q0	Q1	Q2	Q3	Q4	Étendue	Inter quartile
[RESULTAT]	Ordonnée	Perdu	Perdu	Gagné	Gagné (bonus)	Gagné (bonus)	<NA>	<NA>
[POINTS]	Continue	0	14	21,5	30,25	42	0	16,25
[CARTON]	Discrète	0	0	1	1	3	3	1
[METEO]	Nominale	<ND>	<ND>	<ND>	<ND>	<ND>	<NA>	<NA>
[COMMENTAIRE]	Ordonnée	Ultra nul	Mauvais	Moyen	Moyen	Excellent	<NA>	<NA>

TAB. 3.4.4.a – RCV, les quartiles et étendue

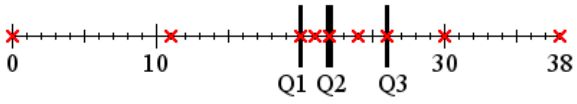
3.4.5 Représentation graphique : la boite à moustache

Dans le cas d'une variable quantitative, la médiane et les quartiles se prêtent à une représentation graphique particulièrement astucieuse, la **boite à moustache** (ou **box plot** en anglais). Cette boite représente d'un seul coup la centralité, la dispersion et met en valeur d'éventuelles points extrêmes à surveiller. Pour la tracer, nous avons besoin des quartiles Q1, Q2, Q3 et de nouveaux concepts propres aux boîtes à moustache : les barrières et les adhérences. Nous les introduirons au fur et à mesure de nos besoins.

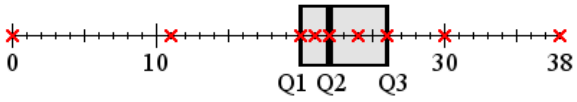
Pour commencer et pour plus de simplicité, abandonnons momentanément nos Villeneuvois pour nous intéresser à leur grand rivaux de toujours : Le Stade de Labarthe. Leur variable [POINT] vaut {0, 11, 20, 21, 22, 24, 26, 30, 38} (9 match, déjà ordonnées).
Pour commencer, considérons une règle graduée allant du minimum 0 au maximum 38 (avec les observations)



Puis traçons un trait vertical large à l'emplacement de la médiane Q2=22, des traits verticaux à l'emplacement des deux quartiles Q1=20 et Q3=26



Ensuite, joignons les quartiles Q1 a Q3 par des segments :



La boîte ainsi obtenue a la propriété de contenir 50% des observations puisqu'elle délimite l'espace inter quartile.

Il nous reste à tracer les moustaches. Pour cela, nous avons besoin des barrières :

Définition 3.4.5.a : Barrière inférieure
La **Barrière inférieure** est une valeur imaginaire en dessous de laquelle une observation est considérée comme extrême. Cette valeur est classiquement fixé à une distance de $1,5 \times EtenduIQ$ au dessous du premier quartile.

$$BarInf = Q1 - 1,5 \times EtenduIQ.$$

Définition 3.4.5.b : Barrière supérieure
La **Barrière supérieure** est une valeur imaginaire au dessus de laquelle une observation est considérée comme extrême. Cette valeur est classiquement fixée à une distance de $1,5 \times EtenduIQ$ au dessus du troisième quartile.

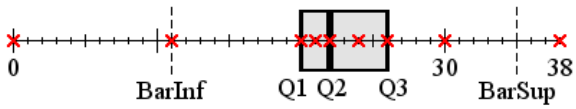
$$BarSup = Q3 + 1,5 \times EtenduIQ.$$

Définition 3.4.5.c : points extrêmes
Les **Points extrêmes** sont les observations dont les valeurs sont strictement inférieure à la barrière inférieure ou strictement supérieure à la barrière supérieure.

$$Extreme = \{Obs < BarInf\} \cup \{Obs > BarSup\}$$

Les barrières sont donc :

- Inférieure : $BarInf = Q1 - 1,5 \times EtenduIQ = 20 - 1,5 \times 6 = 11$
- Supérieure : $BarSup = Q3 + 1,5 \times EtenduIQ = 26 + 1,5 \times 6 = 35$.



Mais ces barrières ne sont là que pour nous permettre de placer les valeurs adhérentes :

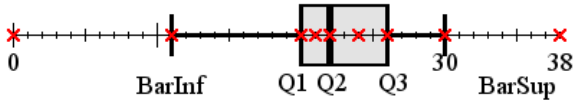
Définition 3.4.5.d : Valeur adhérente inférieure
La **Valeur adhérente inférieure** est la plus petite des observations non extrêmes.

$$ValAdhInf = \min\{Obs \geq BarInf\}$$

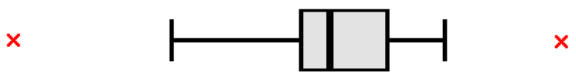
Définition 3.4.5.e : Valeur adhérente supérieure
La **Valeur adhérente supérieure** est la plus grande des observations non extrêmes.

$$ValAdhSup = \max\{Obs \leq BarSup\}$$

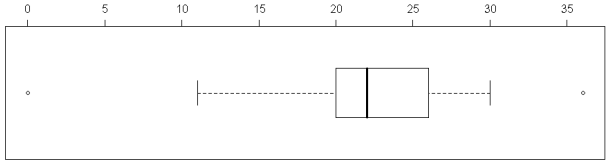
Graphiquement, la valeur adhérente supérieure est donc l'observation la plus proche de la barrière supérieure tout en étant située sur sa gauche (soit 30 sur notre exemple), la valeur adhérente inférieure est l'observation la plus proche de la barrière inférieure et située sur sa droite (soit 11, l'égalité étant accepté). Les "moustaches" sont des traits verticaux situés sur les valeur adhérentes et reliés aux quartiles par une ligne :



Il ne nous reste plus qu'à effacer les traits de construction et la boite à moustache est maintenant terminée :



La même chose faite par le logiciel "R" :



Les boîtes à moustaches sont en particulier utiles pour comparer rapidement les variables de différents groupes. Par exemple, pour la coupe du Comminges, les boîtes à moustaches des scores des différents club nous permettent de voir les clubs réguliers (comme La Serre, abonné aux résultats médiocre, Valentine et Labarthe, pratiquement toujours bon) des clubs irréguliers (comme Villeneuve, Montréjeau et Pointis).

3.4.6 La variance

Les quartiles permettent de mesurer la dispersion. Mais ils souffrent des mêmes défauts que la médiane, en particulier de la non suffisance. D'où le besoin d'un autre indice de dispersion qui serait suffisant.

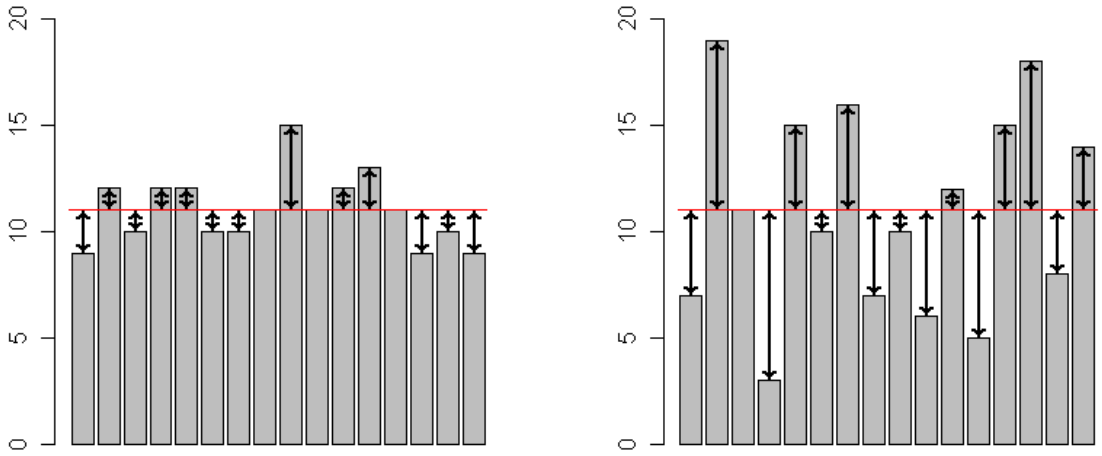
3.4.6.1 L'écart moyen

Le problème est de trouver un indice qui serait grand lorsque les observations sont dispersées, petit quand elles sont regroupées (c'est la définition même de l'indice de dispersion) et qui prennent en compte toutes les valeurs. La première idée qui peut venir à l'esprit est de considérer l'écart entre chaque observation et la moyenne.

Définition 3.4.6.a : Écarts
Étant donnée une variable `[X]` et une observation x_i, l'**Écart à la moyenne** de x_i est la valeur de l'observation *moins* la moyenne des observations.

$$Ecart(x_i) = x_i - \overline{X}$$

Sur l'exemple des notes :



TAB. 3.4.6.a – Contrôle continu, écart a la moyenne

Pour la semaine 2 :

- Les notes sont {9,12,10,12,12,10,10,11,15,11,12,13,11,9,10,9}
- La moyenne vaut 11
- Les écarts sont {-2,+1,-1,+1,+1,-1,-1,0,+4,0,+1,+2,0,-2,-1,-2}

Pour la semaine 3 :

- Les notes sont {7,19,11,3,15,10,16,7,10,6,12,5,15,18,8,14}
- La moyenne vaut 11
- Les écarts sont {-4,+8,0,-8,+4,-1,+5,-4,-1,-5,+1,-6,+4,+7,-3,+3}

Notre objectif est de construire un indice de dispersion. Comment les écarts peuvent-ils être utilisé dans ce sens ? La première idée est d'en faire la moyenne. On obtient

- Semaine 2 : $\frac{-2+1-1+1+1-1-1+0+4+0+1+2+0-2-1-2}{16} = 0$
- Semaine 3 : $\frac{-4+8+0-8+4-1+5-4-1-5+1-6+4+7-3+3}{16} = 0$

Notre tentative d'indice donne un résultat toujours nul ! Cela est dû au fait que les écarts positifs et négatifs se compensent, leur somme est donc nulle. Ce calcul n'aboutit donc pas à un indice de centralité.

Pourtant, l'idée de base était bonne et en prenant, non pas les écarts, mais la *valeur absolue des écarts*, on obtient effectivement un indice de dispersion appelé l'écart moyen.

Définition 3.4.6.b : Écart moyen
L'**Écart moyen** d'une variable `[X]` est la moyenne des valeurs absolues de ses écarts.

$$EM(X) = Moy\{|Ecart(x_i)|\} = \frac{\sum |Ecart(x_i)|}{n} = \frac{\sum |x_i - \overline{X}|}{n}$$

Sur notre exemple :

- $EC(Semaine2) = \frac{|-2|+|1|+|-1|+|1|+|1|+|-1|+|-1|+|0|+|4|+|0|+|1|+|2|+|0|+|-2|+|-1|+|-2|}{16} = 1,25$
- $EC(Semaine3) = \frac{|-4|+|8|+|0|+|-8|+|4|+|-1|+|5|+|-4|+|-1|+|-5|+|1|+|-6|+|4|+|7|+|-3|+|3|}{16} = 4$

L'écart moyen indique une dispersion de 1,25 pour la semaine 2 et une dispersion de 4 pour la semaine 3. Son calcul utilise toutes les valeurs (il est donc suffisant), cet indice semble donc convenir. Pourtant, en pratique, il est assez peu utilisé parce qu'il présente un défaut majeur, invisible aux yeux du profane : pour le calculer, nous avons utilisé des additions, une division et des valeurs absolues. Or, la fonction valeur absolue n'est pas une "bonne fonction"[11] Concrètement, cela rend l'écart moyen difficile à manipuler. D'où le besoin d'un autre indice.

3.4.6.2 La variance

Notre première idée était de considérer les écarts, la valeur absolue n'a été utilisée que pour supprimer les signes négatifs. Il existe une autre fonction permettant de supprimer des signes négatifs, c'est la fonction carré. La fonction carrée est une bonne fonction, dérivable et d'utilisation facile. En calculant la moyenne des carrés des écarts, on obtient un autre indice de dispersion appelé la variance :

Définition 3.4.6.c : Variance
La **Variance** d'une variable `[X]` est la moyenne des carrés de ses écarts.

$$V(X) = Moy\{Carre(Ecart(x_i))\} = \frac{\sum (Ecart(x_i))^2}{n} = \frac{\sum (x_i - \overline{X})^2}{n}$$

Sur notre exemple :

- $V(Semaine2) = \frac{(-2)^2+1^2+(-1)^2+1^2+1^2+(-1)^2+(-1)^2+0^2+4^2+0^2+1^2+2^2+0^2+(-2)^2+(-1)^2+(-2)^2}{16} = 2,5$
- $V(Semaine3) = \frac{(-4)^2+8^2+0^2+(-8)^2+4^2+(-1)^2+5^2+(-4)^2+(-1)^2+(-5)^2+1^2+(-6)^2+4^2+7^2+(-3)^2+3^2}{16} = 21,75$

La variance est donc un indice petit pour la semaine 2, grand pour la semaine 3, suffisant et n'utilisant que des fonctions simples (addition, carré et division). C'est un bon indice de dispersion. Et pourtant...

3.4.7 L'écart type

Dans leur grande majorité, les variables quantitatives sur lesquelles les chercheurs travaillent mesurent une quantité et sont donc dotées d'une unité : les âges de Connors sont en années, les rats traversent des labyrinthes en secondes, les températures de Chuine sont en degré... Lorsque l'on travaille sur des données possédant une unité, on doit respecter certaines règles. En majorité, ces règles sont tellement naturelles qu'on les applique intuitivement. D'autres sont moins évidentes. Pour mémoire :

[11] La notion de "bonne fonction" ou de fonction "mathématiquement sympathique" peut surprendre. Pourtant, il existe en mathématique des bonnes et des mauvaises fonctions. En particulier, une bonne fonction est une fonction qui se laisse facilement manipuler : on peut calculer ses valeurs, (ça n'est pas le cas de toutes les fonctions, il existe des fonctions dont on ne peut pas calculer les valeurs), lui additionner d'autres fonctions, la dériver... La valeur absolue n'entre pas dans ce cadre car elle n'est pas dérivable. C'est donc une fonction que l'on préfère éviter.

- **R1 :** On peut additionner ou soustraire une variable qui a une unité U a une autre variable qui à la même unité U, le résultat est mesuré en U : j'ai couru $1000m$ ce matin et $500m$ cet après midi, soit un total de $1000 + 500 = 1500m$ aujourd'hui.
- **R2 :** On peut multiplier une variable qui a une unité U par un nombre (sans unité), le résultat est mesuré en U : j'ai couru $1000m$ et j'ai recommencé 3 fois, soit $3 \times 1000 = 3000m$
- **R3 :** On peut diviser une variable qui a une unité U par un nombre (sans unité), le résultat est mesuré en U : j'ai couru $2000m$, la moitié le matin et l'autre moitié l'après midi. J'ai donc couru $\frac{2000}{2}m$ ce matin.
- **R4 :** On peut multiplier une variable qui a une unité U par une autre variable qui a une unité U, le résultat est mesuré en U^2 : ce terrain à une longueur de $50m$ et une largeur de $30m$, sa superficie est donc de $1500m^2$
- **R5 :** Dans certain cas, on peut considérer la racine d'une variable. Si la variable a pour unité U^2, sa racine a pour unité U : cette salle est carrée, elle fait $100m^2$, son coté mesure donc $\sqrt{100} = 10m$
- On ne peut pas additionner deux variables dont les unités sont différentes : ce matin, j'ai couru $500m$ sous une température de 17° et j'ai bu $1L$ d'eau...

Fort de ces règles, qu'elle est l'unité de la moyenne ? Une moyenne se calcule en additionnant des variables (R1, pas de changement d'unité) puis en divisant par un nombre (R3, toujours pas de changement d'unité). La moyenne conserve dont l'unité de la variable. Quelle est l'unité de la variance ? Une variance se calcule à partir des écarts (R1, pas de changement d'unité), en les élevant au carré (R4, donc l'unité passe au carré), en additionnant (R1, pas de changement, l'unité reste au carré) puis en divisant par un nombre (R2, pas de changement). L'unité de la variance est donc le carré de celle de la variable. Les températures de Chuine sont en degré, la moyenne des températures est en degré, la variance des températures est en *degré au carré*. Même si cela choque un peu le bon sens, mathématiquement, cela ne pose aucun problème. Ce qui en pose plus, c'est que la moyenne et la variance ne sont pas évaluées dans la même unité. On ne va donc pas pouvoir additionner la moyenne et la variance. On ne va pas plus pouvoir les comparer ou plus simplement les représenter graphiquement sur le même schéma (comme on avait pu faire pour la médiane et les quartiles). La variance est donc un indice de centralité qui a beaucoup de qualités, mais son unité pose un problème. D'où l'introduction d'un nouvel indice, l'écart type :

Définition 3.4.7.a : Écart type
L'**Écart type** d'une variable `[X]` est la racine carré de sa variance.

$$s(X) = \sqrt{V(X)} = \sqrt{\frac{\sum (Ecart(x_i))^2}{n}} = \sqrt{\frac{\sum (x_i - \overline{X})^2}{n}}$$

L'écart type hérite de tous les avantages de la variance et son unité est la racine carré de celle de la variance, c'est à dire la même que celle de la moyenne et des observations. L'écart type est donc un indice avec beaucoup de qualité...

3.4.8 Variance et écart type corrigé

La variance et l'écart type sont à priori des bons candidats dans la course au meilleur indice de dispersion : ils sont cohérents, suffisants (en contre-partie non robuste) et efficaces. Pourtant, ils présentent une faille (et de taille) : ils sont biaisés. Nous présentons en annexe un exemple qui donne une intuition de ce biais. Concrètement, la variance évaluée sur un échantillon est une sous estimation de la variance de la population. Les mathématiciens sont allés plus loin dans leur analyse, ils ont quantifié cette sous estimation : si l'échantillon est de taille n, la variance de la population est en réalité $\frac{n}{n-1}$ fois plus grande que celle de l'échantillon. Pour connaître la véritable variance de la population, il faut donc corriger la variance en la multipliant par un facteur $\frac{n}{n-1}$. Ce nouvel indice est appelé variance corrigée :

Définition 3.4.8.a : Variance corrigée
La **Variance corrigée** d'une variable `[X]` mesuré sur un échantillon de n individus est la variance multiplié par $\frac{n}{n-1}$

$$VC(X) = V(X) \times \frac{n}{n-1} = \frac{\sum (Ecart(x_i))^2}{n-1} = \frac{\sum (x_i - \overline{X})^2}{n-1}$$

De la même manière, l'écart type étant la racine de la variance, l'écart type corrigé est la racine de la variance corrigé :

Définition 3.4.8.b : Écart type corrigé
L'**écart type corrigé** d'une variable `[X]` est la racine carré de sa variance corrigé.

$$sc(X) = Racine(VC(X)) = \sqrt{\frac{\sum(Ecart(x_i))^2}{n-1}} = \sqrt{\frac{\sum(x_i - \overline{X})^2}{n-1}}$$

En pratique, lorsque l'on travaille sur un échantillon de grande taille, $\frac{n}{n-1}$ est pratiquement égal à un et la variance est pratiquement la même que la variance corrigée. Par contre, sur des échantillons de petite taille, la différence peu être plus sensible.

3.4.9 Récapitulatif

Nous pouvons maintenant présenter un tableau complet résumant les variables et les indices de dispersion :

Variable	Nature	Q0	Q1	Q2	Q3	Q4	Étendue	Inter quartile	Variance	Variance corrigée	Écart type	Écart type corrigé
[Resultat]	Ordonnée	Perdu	Perdu	Gagné	Bonus	Bonus	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
[Points]	Continue	0	14	21,5	30,25	42	0	16,25	143,03	152,56	11,96	12,35
[Cartom]	Discrète	0	0	1	1	3	3	1	0,84	0,90	0,92	0,95
[Meteo]	Nominale	<ND>	<ND>	<ND>	<ND>	<ND>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
[Commentaire]	Ordonnée	Ultra nul	Mauvais	Moyen	Moyen	Excellent	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>

Tab. 3.4.9.a – RCV, indices de dispersion

3.4.10 Pour finir : le Best Dispersion Award...

Nous ne reprendrons pas ici toutes les qualités des indices de dispersion [[[principalement parce que je suis bien incapable de calculer la convergence du mode, ou de savoir si l'étendue est biaisé... A voir pour l'an prochain]]] Globalement, les quartiles ont les même avantages et inconvénients que la médiane, l'écart type corrigé a les mêmes avantages et inconvénients que la moyenne. N'importe comment, le calcul de la variance est intimement lié à la moyenne, le calcul des quartiles est lié à la médiane. On ne peut donc pas choisir un indice de dispersion indépendamment de l'indice de centralité. Soit on utilise le bloc médiane-quartile, soit le bloc moyenne-écart type.

En pratique, le bloc moyenne-écart type est préféré, principalement pour sa suffisante et pour son efficacité. Mais si l'on soupçonne les données de contenir des valeurs aberrantes indécelables, il sera plus sage de travailler avec la médiane et les quartiles. Enfin, dans un cas comme dans l'autre, une boîte à moustaches est un bon complément.

3.5 Analyse univarié : bilan

Dans tout ce qui précède, nous avons présenté une analyse univariée exhaustive. En pratique, l'analyse univariée est beaucoup plus courte. Nous vous donnons donc ici un exemple d'analyse univariée, en premier lieu sur un exemple que vous connaissez bien, puis sur un exemple réel.

3.5.1 Le Racing Club de Villeneuve

En 2004, comme tous les ans, le Racing Club de Villeneuve de Rivière a participé à la coupe du Comminges. Voilà le tableau résumant la saison.

[Match]	[Resultat]	[Points]	[Carton]	[Meteo]	[Commentaire]
Match	Resultat	Points	Carton	Meteo	Commentaire
1	Perdu	0	0	Soleil	Ultra nul
2	Gagné (bonus)	124	1	Soleil	Bon
3	Perdu	15	1	Nuageux	Mauvais
4	Nul	11		Soleil	Mauvais
5	Perdu	10	1	Soleil	Ultra nul
6	Perdu	13	2	Soleil	
7	Perdu	6	1	Soleil	Ultra nul
8	Perdu	22	0	Nuageux	Moyen
9	Perdu	11	1	Pluie	Mauvais
10	Gagné (bonus)	35	0	Soleil	Bon
11	Nul	17	2	Soleil	
12	Nul	10	3	Soleil	Mauvais
13	Perdu	3	0	Soleil	Ultra nul
14	Perdu	17	2	Soleil	Moyen
15	Perdu	14	1	Soleil	Mauvais
16	Perdu	0	8	Soleil	Ultra nul

TAB. 3.5.1.a – RCV2004, données brutes

Les variables sont :

Variable	Nature	Modalités
[Resultat]	Qualitative ordonnée	Perdu Nul Gagné Gagné (bonus)
[Points]	Quantitative continue	$[0\,;+\infty]$
[Carton]	Quantitative discrète	$0\,;1\,;2\,;\ldots$
[Meteo]	Qualitative nominale	Soleil Pluie Nuageux Neige
[Commentaire]	Qualitative ordonnée	Très mauvais Mauvais Moyen Bon Excellent

TAB. 3.5.1.b – RCV, liste des variables

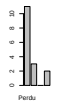
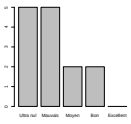
Le résultat de l'analyse univariée est séparé en plusieurs tableaux, selon la nature des variables :

Variable	Nature	Effectif	Mode	NA
[Meteo]	Nominale	Soleil : 13 Pluie : 1 Nuageux : 2	 Soleil	0

TAB. 3.5.1.c – RCV04, variable nominale

Petit détail : par rapport à la saison 2003, la variable météo a une modalité de moins (Neige). Dans notre cas, [Meteo] n'est pas très précise. Pourrait-elle prendre des modalités comme brouillard, verglas, vent ? Rien ne le précise dans cette étude.

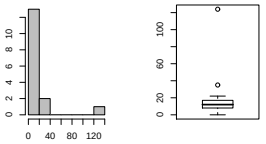
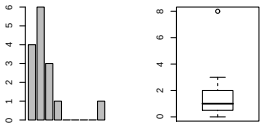
A part ça, rien de particulier n'émerge de ce tableau (si ce n'est que tout de même, il fait beau dans le sud-ouest...)

Variable	Nature	Effectifs	Q0	Q1	Q2	Q3	Q4	NA
[Resultat]	Ordonnée	Perdu : 11 Nul : 3 Gagné : 0 Bonus : 2 	Perdu	Perdu	Perdu	Nul	Bonus	0
[Commentaire]	Ordonnée	Ultra nul : 5 Mauvais : 5 Moyen : 2 Bon : 2 Excellent : 0 	Ultra Nul	Ultra Nul	Mauvais	Moyen	Bon	2

Tab. 3.5.1.d – RCV, indices de dispersion

Ce tableau est nettement plus informatif : la médiane de [RÉSULTAT]est (Perdu) : le RCV a perdu au moins 50% de ses match. Pire le troisième quartile vaut (Nul). Le RCV a gagné moins de 25% de ses match. La saison n'a pas été très bonne...

Les commentaires de l'entraîneur confirment.

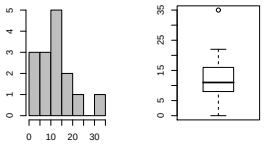
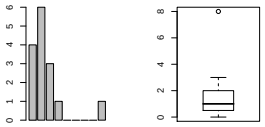
Variable	Nature	Graphique	Q0	Q1	Q2	Q3	Q4	Moyenne	Écart type	NA
[Points]	Continue		0	9	12	17	124	19,25	28,30	0
[Cartons]	Discrète		0	0	1	2	8	1,53	1,93	1

Tab. 3.5.1.e – RCV 2004, variables quantitatives

Deux éléments attirent notre attention : le maximum de [Points] est (124). C'est un résultat très peu probable, sûrement une valeur aberrante. Suite à ce que nous avons publié sur lui, l'entraîneur refuse de nous recevoir. Nous arrivons tout de même, après enquête, à déterminer qu'aucune des autres équipes engagées dans la compétition n'a marqué plus de 60 points sur un match. Nous décidons donc de considérer (124) comme une valeur aberrante.

Un autre problème est posé par la variable [Carton]. Son maximum est 8. Là encore, cela semble assez irréaliste, mais après enquête, personne peut nous affirmer formellement que ça n'est pas arrivé (et nous obtenons la certitude qu'il y a eu dans un autre club 5 cartons). Nous décidons donc de conserver cette donnée. En cas d'analyse plus poussée, nous privilégierons l'utilisation de la médiane (bonne résistance aux valeurs aberrante) à la moyenne.

D'où, après correction, le tableau devient :

Variable	Nature	Graphique	Q0	Q1	Q2	Q3	Q4	Moyenne	Écart type	NA
[Points]	Continue		0	8	11	16	35	12,27	8,90	1
[Carton]	Discrète		0	0	1	2	8	1,53	1,93	1

TAB. 3.5.1.f – RCV 2004, variables quantitatives

On remarque que la suppression de la valeur aberrante n'a pratiquement pas d'impact sur la médiane. Par contre, la moyenne pas de 19,25 à 12,27, ce qui constitue tout de même une variation importante...

Chapitre 4

Approche intuitive des propriétés des indices

Nous allons maintenant donner une approche intuitive de la manière dont sont vérifiés quelques propriétés des indices énoncés section 2.1.1. Naturellement, les minis exemples que nous allons vous présenter n'ont pas la prétention de prouver quoi que ce soit, ils servent juste de support pour approfondir le concept.

Petit rappel : dans une étude réelle, nous disposons d'une population très grande. Sur cette population, nous aimerions connaître la valeur d'un indice. Cette valeur est appelée *le score vraie*). Mais comme la population est trop grande, nous ne pouvons pas calculer le score vraie. Nous nous contentons donc de calculer l'indice sur un échantillon (*valeur estimée*), puis nous généralisons la valeur estimée à la population.

4.1 Biais

Le problème est de savoir si la valeur estimée est une bonne approximation de la valeur vraie.
Pour vérifier cela, nous allons considérer une population fictive composée de 4 (!) individus I1, I2, I3 et I4, une unique variable `[X]` prenant pour valeur $\{1\,;3\,;4\,;6\}$. Ensuite, nous allons extraire de cette population des échantillons de taille 2. Pour une population de taille 4, il existe 6 échantillons possibles : $e_{1.2} = \{I1; I2\}$, $e_{1.3} = \{I1; I3\}$, $e_{1.4} = \{I1; I4\}$, $e_{2.3} = \{I2; I3\}$, $e_{2.4} = \{I2; I4\}$ et $e_{3.4} = \{I3; I4\}$. Pour voir si un indice est biaisé, nous allons calculer la valeur vrai de l'indice (donc la valeur de l'indice sur la population) et la valeur estimée de l'indice sur chacun des échantillons. La moyenne des valeurs estimés est appelé espérance des estimations (c'est la valeur centrale des valeurs estimés ; elle est appelé espérance parce que quand on prend un échantillon au hasard, c'est la valeur la plus probable). Si la valeur vraie et l'espérance des estimations sont égales, l'indice est non biaisé. Sinon, l'indice est biaisé.

Moyenne

- La moyenne de la population est $\overline{X} = \frac{1+3+4+6}{4} = 3,5$.
- Les moyennes des échantillons sont $\overline{e_{1.2}} = 2$, $\overline{e_{1.3}} = 2,5$, $\overline{e_{1.4}} = 3,5$, $\overline{e_{2.3}} = 3,5$, $\overline{e_{2.4}} = 4,5$ et $\overline{e_{3.4}} = 5$.
- L'espérance des moyennes estimées est $\frac{2+2,5+3,5+3,5+4,5+5}{6} = 3,5$
- l'espérance étant égale au score vrai, la moyenne est non biaisée.

Variance

- La variance de la population est $V(X) = \frac{((1-3,5)^2+(3-3,5)^2+(4-3,5)^2+(6-3,5)^2}{4} = 3,25$.
- Les variances des échantillons sont $V(e_{1.2}) = 1$, $V(e_{1.3}) = 2,25$, $V(e_{1.4}) = 6,25$, $V(e_{2.3}) = 0,25$, $V(e_{2.4}) = 2,25$ et $V(e_{3.4}) = 1$.
- L'espérance des variances estimées est $\frac{1+2,25+6,25+0,25+2,25+1}{6} = 2,17$
- l'espérance étant différente du score vrai, la variance est biaisée.

Variance corrigée

- La variance corrigé de la population est $V(X) = \frac{((1-3,5)^2+(3-3,5)^2+(4-3,5)^2+(6-3,5)^2}{3} = 4,33$.

- Les variances corrigées des échantillons sont $V(e_{1.2}) = 2$, $V(e_{1.3}) = 4,5$, $V(e_{1.4}) = 12,5$, $V(e_{2.3}) = 0,5$, $V(e_{2.4}) = 4,5$ et $V(e_{3.4}) = 2$.
- L'espérance des variances corrigées estimées est $\frac{2+4,5+12,5+0,5+4,5+2}{6} = 4,33$
- l'espérance étant égale au score vrai, la variance corrigée est non biaisée.

4.2 Éfficacité

Pour connaître l'efficacité d'un indice, nous allons maintenant considérer des échantillons de tailles variables toujours en considérant notre population fictive : les échantillons de taille 1, les échantillons de taille 2 et enfin les échantillons de taille 3. Selon la théorie, plus un échantillon est important, plus il donne une bonne approximation du score vrai. Les échantillons de taille 3 devraient donc être meilleurs que ceux de taille 2, eux même meilleurs que ceux de taille 1. Comment mesurer "être meilleur" ? En utilisant la variance des estimations. L'idée est que si les estimations sont bonnes, elles ne devraient pas s'écarter du score vrai et donc la variance des estimations devrait être petite. Vérifions :

- Les échantillons de taille 1 : $\overline{e_1} = 1$, $\overline{e_2} = 3$, $\overline{e_3} = 4$ et $\overline{e_4} = 6$. Soit une variance corrigé $V_{Taille1} = 4,33$
- Les échantillons de taille 2 : $\overline{e_{1.2}} = 2$, $\overline{e_{1.3}} = 2,5$, $\overline{e_{1.4}} = 3,5$, $\overline{e_{2.3}} = 3,5$, $\overline{e_{2.4}} = 4,5$ et $\overline{e_{3.4}} = 5$. Soit une variance corrigé $V_{Taille2} = 1,33$
- Les échantillons de taille 3 : $\overline{e_{1.2.3}} = 2,67$, $\overline{e_{1.2.4}} = 3,33$, $\overline{e_{1.3.4}} = 3,67$ et $\overline{e_{2.3.4}} = 4,33$, soit une variance corrigé $V_{Taille3} = 0,48$

Sur notre exemple, plus l'échantillon est de grande taille, plus la variance des estimations est petite (et donc moins on a de chance que notre estimation soit loin du score vrai).

Table des figures

Liste des tableaux

Index

Bibliographie

[Chuine04] I. Chuine, P. Yiou, N. Viovy, B. Seguin, V. Daux, and E. Le Roy Ladurie. *Historical phenology : grape ripening as a past climate indicator.* Nature, 432(7015) :289–90, Novembre 2004.

[Falissard01] B. Falissard. *Mesurer la subjectivité en santé.* Masson, Juillet 2001.

[Harris72] M. B. Harris. *The effects of performing one altruistic act one the likelihood of performing another.* Journal of Social Psychology, 88 :65–73, 1972.

[Reynaud02] M. Reynaud, P. Le Breton, B. Gilot, F. Vervialle, and B. Falissard. *Alcohol is the main factor in excess traffic accident fatalities in france.* Alcoholism, clinical and experimental research, 26(12) :1833–9, Décembre 2002.