Cours 1 : La statistique et les statistiques

Table des matières

Lectures

Obligatoire : Lexique disponible sur le site web.

Suggérée : Howell, chapitre 1 sauf 1.4.

Objectifs

Être en mesure de faire la différence entre population et échantillon, entre données brutes et statistiques (descriptive); de savoir quel type d'échelle a été utilisé pour mesurer les données brutes; de pouvoir faire et lire un graphique des fréquences; de repérer des données abberantes sur un graphique des fréquences.

Section 1. La physique : Une vraie science?

Un physicien (resté anonyme) rapportait qu'« une expérience [de physique] est réussie si je conserve 50% des données brutes ». Cette attitude qui consiste à ne conserver que les données qui font l'affaire du chercheur est très questionnable, et suggère que les physiciens ont une connaissance très superficielle de ce qu'est un processus d'échantillonnage. En psychologie expérimentale, nous ne pouvons pas être aussi naïfs: les observations sont tellement variables d'un individu à l'autre que si on ne gardait que les observations qui répondent à nos attentes, nous éliminerions plus de 90% des données! La solution passe par une meilleure connaissance des principes statistiques pour "faire parler" les données. Un économiste (qui veut rester anonyme) avait coutume de dire "Si on torture les données suffisamment longtemps, elles finiront par se rendre".

Section 2. Science et hasard

La science peut être définie comme l'effort systématique de connaître, de comprendre, et de savoir.

Le but de la recherche scientifique est d'accumuler des observations sur les phénomènes naturels, de les analyser et de les interpréter afin de mieux comprendre leurs causes et conséquences. Les chercheurs mesurent les phénomènes naturels afin de les représenter sous forme numérique et de traiter cette information de façon mathématique.

La traduction d'une observation en une quantité numérique est problématique pour toutes les branches de la science. En psychologie, cette opération peut s'avérer encore plus compliquée puisque l'étude porte sur les comportements et les processus psychiques. Il est cependant important de surmonter cette difficulté étant donnée les multiples avantages qu'offre une formulation en terme mathématique. Entre autre, le langage mathématique est préférable aux langues naturelles pour a) sa précision dans les valeurs et dans les procédures à suivre, b) son universalité.

La science cherche à énoncer les connaissances en termes de lois constantes et générales. En physique et en chimie, il existe un grand nombre de loi. En psychologie, il en existe quelques-uns unes: la loi de Weber sur la perception des différences entre deux stimuli; la loi de Pfitz sur la fréquence des mots et leur nombre de signification; la loi de l'apprentissage décrivant l'amélioration des performances avec la pratique; etc.

Pour arriver à formuler une loi, une seule observation ne suffit pas. La formulation d'une loi requiert une multitude d'observations ou de mesures. Or, la principale caractéristique d'une suite d'observation –surtout chez l'humain- est justement caractérisée par le contraire d'une loi : une variabilité importante dans les mesures.

Cette variabilité provient de trois sources : Deux objets d'études (i.e. deux individus en psychologie) ne sont jamais tous à fait identiques (variabilité inter-sujet). De plus, puisque les personnes sont en perpétuel changement, la mesure d'un même sujet deux fois de suite risque fort d'être différente (variabilité intra-sujet). Finalement, comme une mesure implique un instrument de mesure plus ou moins fidèle et une échelle plus ou moins précise, il est

pratiquement impossible d'éviter des erreurs de mesure. Les statistiques permettent de passer outre le bruit (la variabilité) dans les mesures pour permettre d'énoncer des lois générales.

Section 3. Statistiques descriptives vs. inductives

Les statistiques représentent en fait deux branches, souvent complémentaires. D'une part, il y a les statistiques descriptives. Le but des statistiques descriptives est d'offrir des méthodes pour synthétiser l'information obtenue par un échantillon. Le cours 2 présente plus de détails. D'autre part, il y a les statistiques inductives. Les statistiques inductives (aussi appelée l'inférence statistique) chercher à inférer la valeur d'un paramètre dans la population entière étant donné une ou quelques statistiques tirées d'un échantillon limité. Les cours 4 et suivants présentent des méthodes d'inférences statistiques.

Section 4. Échantillonnage

4.1. Définition

Méthodes permettant de collecter un échantillon. L'échantillon se doit d'être représentatif de la population dont il est extrait. La méthode scientifique précise différents contrôles pour que l'échantillon soit valide (validité externe, i. e. représentatif et doté d'une validité écologique) et consistant (validité interne, i. e. pas contaminé par les attentes du chercheur, effets consécutifs à plus d'une mesure, perte des sujets, etc.). Ces points sont discutés plus en détail dans le cours de méthodologie scientifique. Les statistiques sont incapables de vérifier la validité externe d'un échantillon; cependant, elles peuvent donner quelques indices lorsque la validité interne est faible (voir point 4.4).

Un échantillon n'est jamais parfaitement représentatif d'une population. En effet, nous nous fions au hasard pour choisir les individus que nous allons mesurer. Les différences entre la population et l'échantillon résulte de ce qui est appelé l'erreur d'échantillonnage (aussi appelé –un peu à tord- l'erreur expérimentale). Elle reflète les erreurs de mesures (voir ci-après) et le hasard dans le choix de nos observations.

4.2. Échelles de mesures

Une observation (une donnée brute) présentée sous forme numérique est toujours déterminée par trois aspects : Avant tout par l'attribut qui est mesuré. Il importe de toujours garder à l'esprit qu'on ne peut jamais mesurer un objet naturel dans sa totalité, mais seulement un attribut à la fois.

Dans une moindre mesure, il dépend aussi de la précision de l'instrument de mesure et des caractéristiques propres de l'échelle de mesure. Une mesure contient immanquablement une erreur de mesure plus ou moins grande et de cette précision découle le choix de l'échelle de mesure utilisée.

On rencontre quatre types d'échelles : l'échelle nominale, l'échelle ordinale, l'échelle relative, et l'échelle absolue.

a. L'échelle nominale

L'échelle nominale implique un simple groupement des observations en catégories qualitatives identifiées par un symbole (souvent une étiquette, tel « Homme » et « Femme » pour identifier le sexe). La seule opération mathématique possible avec cette échelle est de compter le nombre d'éléments (les effectifs) dans chacune des catégories (parfois nommées des classes), qu'on appelle aussi la fréquence observée ou plus simplement, la fréquence.

b. L'échelle ordinale

L'échelle ordinale est similaire à l'échelle nominale exceptée qu'elle permet d'établir une relation d'ordre entre les éléments d'un ensemble, sans toutefois être capable d'évaluer de façon quantitative la distance qui les sépare. Dans l'exemple précédent, il est impossible de dire si la catégorie « Homme » doit être placée avant la catégorie « Femme ». Un exemple d'échelle ordinale est donnée par les notes scolaires. Clairement, un A vaut mieux qu'un B, qui lui-même est meilleur qu'un C, etc. Cependant, avoir A ne signifie pas que l'étudiant maîtrise deux fois plus la matière que celui qui a un B. Une échelle ordinale représente des rangs. Avec cette échelle de mesure, on peut calculer des fréquences, mais aussi des moyennes et d'autres statistiques. Dans ce contexte, la moyenne doit être comprise comme le rang moyen.

c. L'échelle relative

L'échelle relative (encore appelé l'échelle à intervalles) définit numériquement les intervalles entre les données. Cette échelle possède une unité de mesure arbitraire mais constante. Cependant, le zéro sur ces échelles est défini de façon arbitraire. Un exemple est la température exprimée en celsius. Zéro Celsius est un point arbitraire qui a été choisi par convention. D'ailleurs les échelles Fahrenheit et Celsius n'ont pas le même zéro. Passer de 10 à 15 Celsius demande le même travail (le même nombre de joules) que pour passer de 40 à 45 Celsius. Cependant, cette échelle de mesure ne permet pas d'affirmer que de l'eau à 10 Celsius est deux fois plus chaude que de l'eau à 5 Celsius. Un autre exemple d'échelle relative est l'échelle de Q.I. où le Q.I. moyen est arbitrairement placé à 100.

d. L'échelle absolue

L'échelle absolue (parfois appelée échelle de rapport) implique que la distance entre deux unités est la même tout au long de l'échelle (tout comme dans l'échelle relative) mais aussi que le zéro existe (autrement que par un choix arbitraire). En plus de permettre de quantifier la différence entre deux éléments, elle permet aussi de calculer des rapports entre deux mesures. Par exemple, une distance de 4 mètres est belle et bien le double d'une distance de 2 mètres. Un autre exemple est la température en Kelvin. De l'eau à 300 Kelvin est deux fois plus chaude que de l'eau à 150 Kelvin en ce sens que l'on peut en extraire deux fois plus d'énergie cinétique.

Les quatre types d'échelle ont été présentés dans un ordre ascendant de précision. Dans la suite, nous nommerons échelles de type I les échelles nominale et ordinale et échelles de type II les échelles relative et absolue. Chacune de ces échelles possède les propriétés des échelles qui lui sont inférieures en plus de ses caractéristiques propres. Il est toujours possible

de passer d'une échelle d'un niveau donné à une échelle moins précise; l'inverse n'est cependant pas possible.

4.3. Représentations graphiques des données brutes

La façon la plus précise de regarder vos données brutes est de regarder leur distribution en utilisant un graphique des fréquences (souvent appelé graphique des histogrammes). Pour ce faire, il faut répartir les observations en différentes classes et compter la fréquence des observations correspondant à chacune de ces classes. Nous construisons ainsi une distribution de fréquences.

Si les données sont disposées sur une échelle de type I (nominale ou ordinale), un graphique en histogramme peut directement être obtenu des données brutes. Voir la Figure 1 pour un exemple,

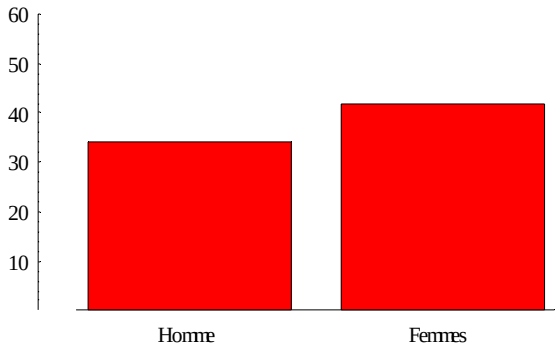

Figure 1 : Exemple de graphe de répartition des données brutes

Pour des échelles plus précises (de type II), où il existe virtuellement une infinité de valeurs possibles, il est nécessaire de regrouper les valeurs obtenues en classes successives (i. e. des intervalles). Par exemple, il n'est pas très intéressant de savoir qu'une personne dans notre échantillon mesurait 1.74824 m. Il est plus intéressant de savoir que 76 d'entre elles avaient une taille entre 1.70 m et 1.80 m. Ce faisant, on se trouve à établir la distribution de nos données brutes qui nous indique alors comment se répartissent nos mesures. Voici un

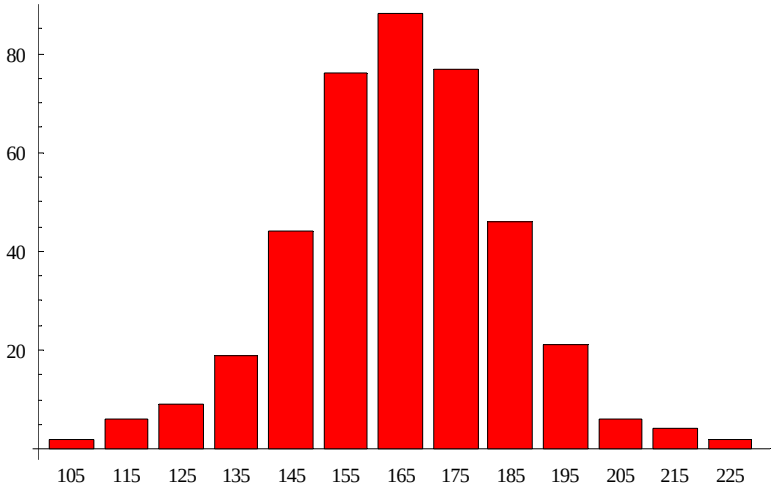

Figure 2 : Exemple de graphe de répartition des tailles d'individus

exemple avec des tailles à la Figure 2.

Pour construire une distribution de fréquences manuellement, il est préférable de commencer par trier les données. Ensuite, il est nécessaire de : a) calculer l'étendue de la série de données et b) de choisir le nombre d'intervalle que nous allons utiliser. Le nombre d'intervalle dépend à la fois de l'étendue des données et du nombre total des données

a. L'étendue d'une série de données

L'étendue est l'écart entre la plus grande et la plus petite donnée brute. En d'autres termes, l'étendue = Max(**X**) – Min(**X**). L'étendue est utile pour graduer l'abscisse du graphique des fréquences qui va suivre.

b. Le nombre d'intervalles

Le nombre d'intervalles (ou encore de classes, de catégories) indique en combien de catégories l'on doit regrouper les données. En règle générale, si le nombre de données brutes est moyen (quelques centaines), il faut subdiviser l'étendue en environ 10 à 20 intervalles égaux (ce nombre peut varier légèrement pour permettre de choisir des classes simples, par exemple, distribuer les données de 10 en 10). Quand le nombre d'observations *n* excède 200, on peut utiliser plus de classes si on le souhaite, pour donner un aspect plus lisse au graphique. Dans tous les cas, la taille de chaque intervalle doit être identique, égale à l'étendue divisée par le nombre de catégories voulues.

c. Compilation des données

Pour chaque intervalle de classe, il faut calculer :

a) Le centre de la classe, i. e. la valeur équidistante aux bornes de chaque classe. Elle constitue la valeur représentative de la classe, et est utilisée pour graduer l'abscisse.

b) Le nombre de données incluses dans cette classe, ce qu'on appelle la fréquence ou encore l'effectif.

c) La fréquence relative (*f*) qui constitue l'effectif de chaque classe exprimé en proportion de l'effectif total.

d) La fréquence cumulative (F), c'est à dire la somme de l'effectif de cette classe et des classes précédentes.

Exemple :

Soit les Q.I suivant, collectés dans une classe de psychologie. X={47, 48, 81, 83, 83, 86, 87, 89, 100, 100, 101, 101, 101, 102, 102, 102, 103, 103, 103, 103, 104, 104, 104, 105, 105, 106, 106, 106, 106, 107, 107, 107, 107, 108, 108, 108, 108, 109, 109, 109, 109, 110, 110, 111, 111, 112, 112, 113, 113, 114, 114, 115, 115, 115, 116, 116, 117, 117, 117, 118, 119, 120, 121, 121, 122, 122, 122, 123, 123, 123, 124, 124, 124, 124, 125, 125, 125, 125, 126, 126, 126, 126, 127, 127, 127, 126, 128, 128, 128, 129, 129, 130, 131, 131, 132, 132, 133, 133, 134, 135, 136, 137, 139, 141, 141, 142, 143, 144, 145, 146, 146, 148, 152, 155, 163, 168, 172}. Ces données ont été triées pour simplifier la tâche.

Puisque les données s'étendent de 40 à 172, nous avons une étendue de 132. Si l'on divise les données par intervalles de 10 (plus facile à travailler), nous aurons 14 classes, ce qui est bien raisonnable.

Intervalle	Centre	*Fréquence*	*Fré. Cum.*	*f*	*F*
40 .. 50	45	2	2	1.7%	1.7%
50 .. 60	55	0	2	0.0%	1.7%
60 .. 70	65	0	2	0.0%	1.7%
70 .. 80	75	0	2	0.0%	1.7%
80 .. 90	85	6	8	5.1%	6.8%
90 .. 100	95	2	10	1.7%	8.5%
10 .. 110	105	33	43	28.2%	36.8%
110 .. 120	115	19	62	16.2%	53.0%
120 .. 130	125	30	92	25.6%	78.6%
130 .. 140	135	11	103	9.4%	88.0%
140 .. 150	145	9	112	7.7%	95.7%
150 .. 160	155	2	114	1.7%	97.4%
160 .. 170	165	2	116	1.7%	99.1%
170 .. 180	175	1	117	0.9%	100.0%
(incl.) (excl.)	**Total :**	**117**		**100%**	

Le graphique des fréquences résultant est donné à gauche, et celui des fréquences cumulatives, à droite de la Figure 3.

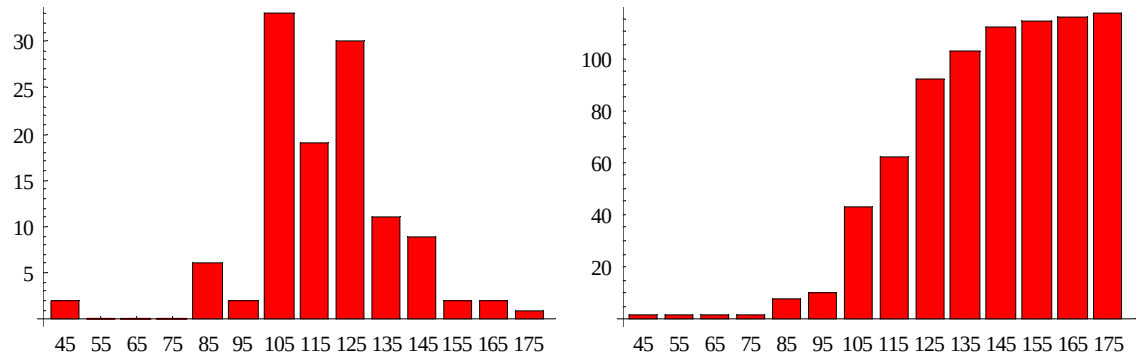

Figure 3 : Graphique des fréquences et graphique des fréquences cumulatives obtenus sur les même données.

Le graphique des fréquences cumulatives doit toujours atteindre la valeur *n*, soit le nombre total de données brutes (117 dans cet exemple).

Pour donner un exemple où le nombre de classe est mal choisi, j'ai refait à la Figure 4 le même graphique, mais cette fois avec respectivement trois et 140 catégories. Le premier cas

semble indiquer que le Q.I est parfaitement symétrique autour de 115, ce qui n'est pas le cas; le second graphique est rempli de catégories vides, donnant un aspect très aléatoire à notre échantillon.

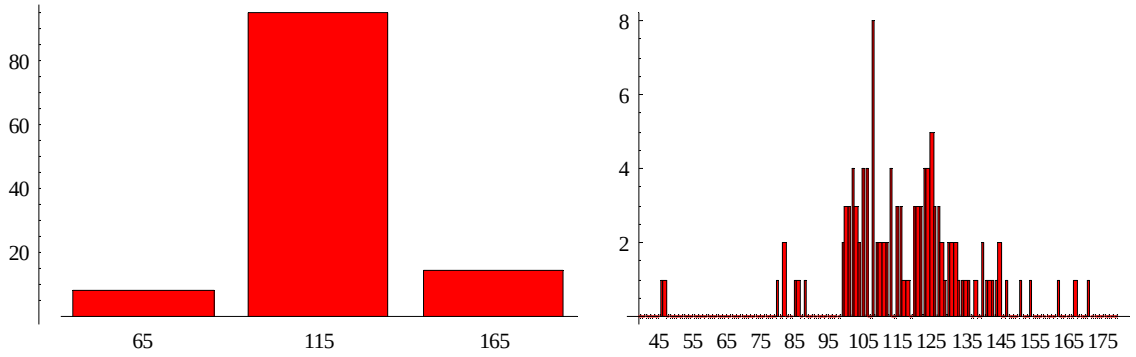

Figure 4 : Graphiques des fréquences pour lesquels le nombre de classes a été mal choisi

4.4. Validation de l'échantillon

La distribution des données n'est pas souvent présentée en psychologie car peu de chercheurs ont des hypothèses sur celle-ci. Cependant, il est néanmoins indispensable de faire le graphique des fréquences. En effet, la seule façon empirique de vérifier que l'échantillon a été collecté dans des conditions uniformes est de s'assurer que les observations se distribuent de façon continue. Bien qu'il s'agisse d'une notion floue, deux points devraient attirer votre attention à cette étape : a) la présence de données extrêmes, et b) la présence de multimodalité.

Les données extrêmes ne sont pas forcément incorrectes, mais elles doivent être considérées comme suspectes. Dans l'exemple sur les Q.I., on voit qu'il existe deux données extrêmement faibles par rapport aux autres. Ces données sont facilement visible sur un graphique des fréquences (voir Figure 3) puisqu'elles forment une barre isolée (à droite ou à gauche). Le chercheur qui obtient des données extrêmes doit s'interroger pour savoir si les sujets qui ont donné ces résultats sont bien représentatifs de l'échantillon. Dans le cas des Q.I., c'est clairement suspect puisqu'il est étonnant d'atteindre l'université avec un Q. I. inférieur à 70. Quand un chercheur est confronté à quelques données extrêmes (jamais plus de 2% de ces données), il peut les omettre des analyses subséquentes à la condition de bien préciser dans son rapport qu'il a retiré des données (lesquelles et combien). S'il existe plus de 2% de données extrêmes, le chercheur ne peut pas les éliminer (il élimine une trop grande part de son échantillon), et doit plutôt questionner sa méthode d'échantillonnage (validité interne). Il devrait peut-être considérer des critères de sélection plus sévères ou veiller à ce que la motivation de ses participants soit plus égale.

La multimodalité se voit par la présence de plusieurs pics distincts dans le graphique des fréquences. Par exemple, dans la Figure 3, il y a deux pics très distincts. Une fois encore, la présence de multimodalité n'est pas en soi le signe d'une erreur dans les données. Cependant, le chercheur doit vérifier que ce ne sont pas systématiquement des sujets différents qui se trouvent dans des pics différents. Par exemple, peut-être que le premier pic

est composé principalement d'étudiants de première année alors que le second pic est composé majoritairement d'étudiants de seconde année. Le chercheur doit alors se poser la question s'il ne doit pas séparer son échantillon en deux pour éviter des conclusions qui ne seraient vraies que pour une partie des étudiants de première année et une partie des étudiants de seconde année. Un échantillon non homogène peut créer des problèmes de validité interne plus tard lorsque le chercheur voudra interpréter ses résultats.

Section 5. Conclusion

Exercices

1. Une caractéristique ou un phénomène pouvant prendre différentes valeurs est :
   a) Une constante
   b) Une donnée brute
   c) Une population
   d) Une variable
   e) Un paramètre
2. Une caractéristique de la population qui est mesurable est :
   a) Une constante
   b) Une donnée brute
   c) Une population
   d) Une variable
   e) Un paramètre
3. Un nombre résultant d'une manipulation de données brutes d'un échantillon, en respectant certaines procédures, se nomme :
   a) Une statistique
   b) Une donnée brute
   c) Une variable
   d) Un paramètre
4. La moyenne d'un échantillon est :
   a) Une statistique inductive
   b) Une statistique descriptive
   c) Un paramètre
   d) Une constante
   e) Une donnée brute
5. L'ensemble de tous les adultes ayant le droit de voter est :
   a) Une statistique
   b) Une population
   c) Un échantillon
   d) Une constante
   e) Une donnée brute
6. Le but de la statistique inductive est de faire des inférences sur _____ à partir ___ issu(es) _____.
   a) Un échantillon, d'une population, de cet échantillon
   b) Une statistique, d'un échantillon, de cette population
   c) Une statistique, d'une population, de statistiques
   d) Une population, d'un échantillon, de cette population
   e) Une population, d'un échantillon, de statistiques.
7. Les observations faites sur un échantillon diffèrent les unes des autres parce que :
   a) Les objets observés sont différents
   b) Les instruments utilisés sont imprécis
   c) Les expérimentateurs font des erreurs
   d) Aucune de ces réponses
   e) a, b, et c sont vraies.
8. Suite au recensement d'une communauté urbaine, il est reporté que 53% des familles ont deux enfants ou plus. Il s'agit :
   a) D'un paramètre de la population
   b) D'une statistique de la population
   c) D'un paramètre de l'échantillon
   d) D'une inférence statistique
   e) D'une constante

9. Parmi les expressions suivantes, où avons-nous correctement utilisé une statistique :

   a) Ses statistiques sont 1m75 et 90 kg

   b) En terme de statistique des ventes, le mois dernier, nous avons vendu un manteau de 1000$

   c) À l'Université de Montréal, Jeanne n'est qu'une statistique de plus

   d) Le coût moyen pour instruire un-e étudiant-e est de 10 000$

   e) Toutes ces réponses

10. Laquelle de ces expressions est vraie :

    a) Le but de la science est de trouver des lois générales

    b) Les observations diffèrent les unes des autres

    c) L'infidélité des instruments de mesure est une des sources de variabilité des observations

    d) L'imprécision d'une échelle de mesure est une source de variabilité

    e) Toutes ces réponses.

11. Soit la variable aléatoire **X** comprenant les valeurs { 1, 2, 3, 4, 5, 6, 7, 8, 9, 10}. Quelle est la somme de ces valeurs:

    a) $\sum \mathbf{X}_i$

    b) 10 (1 + 10) / 2

    c) n (n + 1) / 2

    d) 55

    e) Toutes ces réponses

12. « La somme des valeurs de la variable **X** divisée par le nombre de valeurs » est notée:

    a) $\sum$

    b) $\sum n / \mathbf{X}$

    c) $\sum \mathbf{X}_i / n$

    d) Toutes ces réponses

    e) Aucune de ces réponses

13. $\sum \left(\mathbf{X}_i - \sum \mathbf{X}_i / n \right) =$

    a) 0

    b) $\sum \mathbf{X}_i / n$

    c) n

    d) **X**

    e) Impossible à déterminer

14. Déterminer le type d'échelle de ces mesures :

    a) Age : _____

    b) Ethnie : _____

    c) Rang social : _____

    d) Poids : _____

15. Le regroupement d'individus dans des catégories telles que « Faible », « moyen » et « Fort » implique quel type d'échelle?

    a) Nominal

    b) Ordinal

    c) Relative

    d) Absolue

16. L'échelle caractérisée par la classification d'objets dans des catégories où la seule relation possible est celle d'appartenance est appelée :

    a) Nominal

    b) Ordinal

    c) Relative

    d) Absolue

17. La transformation d'une échelle à une autre n'est pas possible dans le cas suivant :

a) De nominale à relative

b) De relative à ordinale

c) D'absolue à nominale

d) D'absolue à relative

18. Lorsque l'on dit « Mario est plus beau que Simon », quel type d'échelle utilise-t-on?

a) Nominal

b) Ordinal

c) Relative

d) Absolue

19. Lequel permet la mesure la plus précise?

a) L'ordre des chevaux à l'arrivé

b) Le nombre d'homme et de femmes en psychologie

c) La température en Celsius

d) La distribution des votes lors d'une élection

e) La distance entre la terre et les planètes du système solaire.

20. Une échelle pour laquelle la distance entre les points est approximative est appelée :

a) Nominale

b) Ordinale

c) Relative

d) Absolue