L'ESCALE
CENTRE D'ÉDUCATION DES ADULTES

COMMISSION SCOLAIRE DE L'AMIANTE
Un partenaire à la réussite !

DOCUMENT DE RÉVISION MAT-4104

ÉLABORÉ PAR RICHARD ROUSSEAU, ENSEIGNANT EN MATHÉMATIQUES,
CENTRE D'ÉDUCATION DES ADULTES L'ESCALE

COMMISSION SCOLAIRE DE L'AMIANTE

MAI 2005

Mathématiques – Statistiques II

MAT-4104-2

Collecte de données :

Il y a trois modes de collecte de données :

- Le recensement : le recensement est une étude statistique qui porte sur toute une population.

- Le sondage : le sondage est une étude statistique qui porte seulement sur une partie d'une population ciblée et grâce aux données recueillies, on tire des conclusions sur la population entière.

- L'enquête : l'enquête est une étude statistique approfondie sur un sujet et nécessitant l'intervention d'experts en la matière.

Exemples :

Nous voulons savoir dans quelle proportion les personnes dans la ville où tu habites préfèrent le Coca-Cola, le Pepsi ou bien le 7-up. Il ne sera pas nécessaire de demander à toute la population de la ville alors c'est le sondage qui sera le plus approprié.

Un nouveau médicament verra le jour mais il faut le tester sur des individus pour connaître les effets secondaires, s'il y en a. L'enquête sera la plus appropriée car un petit échantillon de personnes fera l'affaire et il doit être suivi par des experts en pharmacologie.

Dans ta ville, les élections municipales auront lieu bientôt alors il faudra recenser toute la population de ta ville afin de savoir qui aura le privilège de voter.

Recueillement de données :

Pour recueillir les données, il y a l'entrevue téléphonique, le questionnaire écrit, l'entrevue, l'observation en directe et l'utilisation de moyens électroniques ou mécaniques.

Caractéristiques d'une population ou d'un échantillon:

Un échantillon est représentatif d'une population si toutes les caractéristiques de la population se retrouvent dans l'échantillon ainsi, chaque individu de la population aura la même chance que les autres de faire partie de l'échantillon sinon celui-ci sera biaisé.

Par exemples :

- Un nombre suffisant de personnes compte tenu d'une marge d'erreur.
- Une représentation égale de personnes de sexe masculin et féminin.
- Une représentation proportionnelle des groupes d'âges.

La taille de l'échantillon :

La taille de l'échantillon est très importante : si l'échantillon est trop petit alors il ne sera pas représentatif de la population et faussera les résultats.

Voici un tableau permettant de déterminer la taille d'un échantillon qui correspond à une taille de population donnée.

	Marge d'erreur maximale		
Taille de population	± 1 %	± 3 %	± 5 %
250	244	203	152
1 000	906	516	278
10 000	4899	964	370
100 000	8763	1056	383
500 000	9423	1065	384
5 000 000	9586	1067	384

Source : Sample Size Calculator, The Survey System : Your complete Software Sloution for Suvey Research, Pelalume (Californie, État-unis), Creative Research Systems, 2002, www.wurveysystem.com/sscalc.htm

Par exemple, si nous voulons sonder une population de 100 000 sujets et que nous voulons que 95 % de l'échantillon soit représentatif de la population, alors il faudra un échantillon de 383 individus et si, par contre, nous voulons une marge d'erreur de ±1 %, autrement dit avec une certitude de 99 % que notre échantillon sera représentatif de la population, nous devrons avoir un échantillon de 8 763 individus.

Il y a aussi une formule afin de déterminer la taille de l'échantillon d'une population de plus de 100 000 individus étant donné la population visée et le pourcentage de marge d'erreur désiré.

Formule pour trouver la taille d'un échantillon.

$n = \frac{0{,}9604}{E^2}$ Où **n** est la taille de l'échantillon et ***E*** est la marge d'erreur.

Exemple :

Un député désire faire un sondage pour connaître les intentions de vote des citoyens de la ville où il habite. La ville compte 280 000 habitants et il désire obtenir une marge d'erreur de ± 5 %. Combien de citoyens seront alors interrogés?

Il s'agit de remplacer le *E* dans la formule par 5 %.

Alors $E = 5\ \% = \frac{5}{100} = 0{,}05$ et,

$$n = \frac{0{,}9604}{E^2} \rightarrow n = \frac{0{,}9604}{0{,}05^2} \rightarrow n = \frac{0{,}9604}{0{,}0025} \rightarrow n = 384{,}16$$

Donc, 384 personnes seront interrogées pour que l'échantillon soit représentatif de la population avec une marge d'erreur de ± 5 %. Il est à noter que ce nombre est aussi celui donné par le tableau.

Autre exemple :

Dans une ville de 10 000 personnes, nous désirons prendre un échantillon de 400 personnes et faire un sondage sur le choix d'un candidat lors de la prochaine élection municipale. Quel sera alors le pourcentage d'erreur, autrement dit, est-ce que l'échantillon sera très représentatif de l'ensemble de la population?

Nous avons le n = 400 alors, il s'agit de trouver le *E*.

$\mathbf{n} = \frac{0{,}9604}{E^2}$ donc $400 = \frac{0{,}9604}{E^2}$ et $E^2 = \frac{0{,}9604}{400}$ et $E^2 = 0{,}002401$

Alors $E = \sqrt{0{,}002401}$ donc 4,9%.

L'échantillon sera représentatif de la population à 95,1 %

Le biais, ou ce qui peut fausser les résultats d'une étude statistique.

Les éléments susceptibles de biaiser les résultats d'une étude statistique sont :

- Le choix d'un l'échantillon qui peut ne pas être représentatif de la population : la sélection des individus est mauvaise ou la taille de l'échantillon n'est pas adéquate. La sélection des individus doit être aléatoire et chaque individu doit avoir une chance égale d'être sélectionné. La taille de l'échantillon doit être adéquate par rapport à la taille de la population.

- Le procédé de collecte de données tels de mauvaises questions ou l'utilisation d'expressions difficiles à comprendre.

- Le traitement et l'analyse des données tels des erreurs de mesure et la présentation des résultats.

Interprétation des résultats d'un sondage :

Étant donné les résultats d'un sondage dont on connaît la marge d'erreur, nous allons déterminer dans quel intervalle se situe un résultat selon que l'on tient compte ou non des indécis.

Exemple :

Un sondage sur la préférence des gens pour le Coca-Cola et le Pepsi donne les résultats suivants :

Préférences de goût pour une boisson gazeuse

Breuvage	Fréquence relative
Coca-Cola	35 %
Pepsi	45 %
Indécis	20 %

Avec une marge d'erreur de ± 3 %, 19 fois sur 20

Étant donné une marge d'erreur de ± 3 %, cela signifie que les résultats varieront de plus ou moins 3 %.

Pour le Coca-Cola, nous avons une fréquence de 35 % alors,

35 % – 3 % = 32 % et,
35 % + 3 % = 38 %

Cela signifie que le pourcentage de gens qui ont préféré le Coca-Cola pourrait varier entre 32 % et 38 %.

Pour le Pepsi, nous avons une fréquence de 45 % donc,

45 % – 3 % = 42 %
45 % + 3 % = 48 %

Cela signifie que le pourcentage de gens qui ont préféré le Pepsi pourrait varier entre 42 % et 48 %

<u>À toi maintenant de trouver l'intervalle pour les indécis et demande à ton formateur de vérifier ta réponse.</u>

Autre exemple :

Dans un poste de radio, pendant une heure, l'animateur a demandé à ses auditeurs de choisir pour l'album de tous les temps parmi trois choix. Les résultats sont représentés ci-dessous avec une marge d'erreur de ± 5%.

Album de tous les temps	
Album	Fréquence
Dark Side of the Moon de Pink Floyd	268
Ok Computer de Radiohead	86
The Joshua Tree de U2	69
Indécis	97

En ne tenant pas compte des indécis, dans quel intervalle se situe le disque Ok Computer?

En ne tenant pas compte des indécis, le total de personnes interrogé est de 423 donc, le pourcentage de gens ayant choisit Ok Computer est de : $\frac{86x100}{423} = 20\%$.

Alors, l'intervalle se situe à ± 5%.
20% - 5% = 15%
20% + 5% = 25%.

<u>En tenant compte des indécis, calcule maintenant l'intervalle où se situe le disque Dark Side of the Moon. Demande à ton formateur de vérifier ta réponse.</u>

Mesures de tendance centrale

Les mesures de tendance centrale permettent d'analyser les données recueillies. Les trois mesures de tendance centrale sont : le mode, la moyenne et la médiane.

Le mode

Le mode d'une distribution est la donnée qui apparaît le plus fréquemment dans une distribution. On représente le mode par les lettres **Mo.**

Exemple : Nombre d'animaux domestiques par foyer

Nombre d'animaux	Foyers
0	142
1	98
2	65
3	45
4	3
5	1
Total	354

La fréquence la plus élevée étant 142, le mode de cette distribution est donc Mo = 0. Il y a 142 foyers qui ne possèdent pas d'animaux domestiques.

Autre exemple:

Voici un tableau représentant la masse en kg des élèves de ma classe.

Masse kg	**Fréquence**
[40,50[	8
[50,60[	7
[60,70[	12
[70,80[	4
[80,90[	2

La fréquence la plus élevée étant 12, le mode se situe dans la **classe modale** [60,70[.

Le mode sera la valeur centrale de la classe modale et nous le calculons ainsi:

Mo = $\frac{70+60}{2}$ donc, Mo = 65kg

La moyenne

La moyenne d'une distribution est obtenue en additionnant la somme de toutes les valeurs divisée par le nombre total de données de cette distribution. La moyenne est représentée par le symbole ξ.

La formule est la suivante : $\xi = \frac{\sum x_i}{n}$ où $\mathbf{x_i}$ représente les valeurs des données et **n** le nombre de données. Le symbole Σ signifie la somme des valeurs de x.

En soit, la formule veut dire : $\xi = \frac{x_1 + x_2 + ... + x_n}{n}$

Exemple : Nombre d'animaux domestiques par foyer

Nombre d'animaux	Foyers
0	142
1	98
2	65
3	45
4	3
5	1
Total	354

La formule de la moyenne est : $\xi = \frac{\sum x_i}{n}$

Donc, $\xi = \frac{142+98+65+45+3+1}{6} = \frac{354}{6} = 59$

La moyenne est donc $\xi = 59$

Autre exemple :

Reprenons le tableau représentant la masse des élèves de ma classe.

Masse kg	**Fréquence**
[40,50[	8
[50,60[	7
[60,70[	12
[70,80[	4
[80,90[	2

Quelle est la masse moyenne des élèves de ma classe?

Nous devons trouver en premier le milieu de chacune des classes.

Le milieu de chaque classe se trouve ainsi :

Pour la première classe : $\frac{50+40}{2}$ = 45 et ainsi de suite pour les autres.

Ensuite, il faut multiplier par la fréquence.

Milieu de la classe kg	**Fréquence**	**Fréquence x classe**
45	8	360
55	7	385
65	12	780
75	4	150
85	2	170
Total	33	1845

Et trouver la moyenne : $\xi = \frac{1845}{33}$ = 55,90 kg

La médiane

La médiane d'une distribution est la valeur qui est située au centre de cette distribution lorsque les données sont classées en ordre croissant. La médiane est représentée par les lettres **Md**.

Pour trouver la médiane d'une distribution, il faut :

- Classer les données en ordre croissant;
- Déterminer le rang de la donnée qui se situe au centre de la distribution.

Exemple :

Nombre impair de données.

Pour ses mathématiques 536, Marc a obtenu les résultats suivants:

83%, 77%, 67%, 80%, 71%, 97%, 65%

Quel est la médiane de ses notes de mathématiques?

Lorsque le nombre de données est impair, le rang de la médiane se trouve à l'aide de la formule : $\frac{(n+1)}{2}$ où n représente le nombre de données de la distribution.

Il faut d'abord classer les notes en ordre croissant.

65%, 67%, 71%, 77%, 80%, 83%, 97%

Et déterminer la donnée qui est située au centre de la distribution.

Le rang est $\frac{(n+1)}{2} = \frac{(7+1)}{2} = 4$. Alors la médiane se situe au $4^{\text{ième}}$ rang et elle vaut Md = 77%

65%, 67%, 71%, $\boxed{77\%}$, 80%, 83%, 97%

Nombre pair de données.

Pour ses mathématiques 436, Jean a obtenu les résultats suivants:

86%, 75%, 82%, 68%, 90%, 89%, 70%, 92%

Quel est la médiane de ses notes de mathématiques?

Lorsque le nombre de données est pair, le rang de la médiane se trouve à l'aide de la formule : $\frac{(n)}{2}$ et $\frac{(n)}{2}$ + 1 où n représente toujours le nombre de données de la distribution. Avec la formule, il faut trouver le rang des deux données centrales de la distribution, les additionner et diviser par deux.

Classons d'abord les notes en ordre croissant.

68%, 70%, 75%, 82%, 86%, 89%, 90%, 92%

Les rangs sont : $\frac{n}{2} = \frac{8}{2} = 4$ et $\frac{n}{2} + 1 = \frac{8}{2} + 1 = 4 + 1 = 5$. Alors il faut prendre la $4^{\text{ième}}$ et la $5^{\text{ième}}$ donnée, les additionner et diviser par deux. $\frac{82\% + 86\%}{2} = 84\%$.

Donc, la médiane est 84%.

Exercices :

A) Le tableau suivant donne le nombre d'heures écouté par une clientèle âgée de 18 à 25 ans pour un poste de radio de la région.

Nombre d'heures d'écoute

0	2 4 8 9
1	1 2 2 4 6 9 9 9
2	1 1 2 2 2 5 8 8 9
3	0

Trouve chacun des items suivants et fais corriger tes réponses par ton formateur.

1) la moyenne
2) le mode
3) la médiane

B) Trouve l'étendue, le mode, la moyenne et la médiane de la série suivante et fais corriger par ton formateur.

12, 13, 26, 28, 28, 28, 28, 30, 34, 37, 37, 38, 39, 40, 42, 46, 48, 50

Mesures de dispersion

L'étendue:

L'étendue d'une distribution est la différence entre la donnée la plus grande et la plus petite.

Dans la distribution suivante, l'étendue est 98 – 55 = 43

55 – 58 – 60 – 62 – 72 – 74 – 78 – 78 – 80 – 86 – 89 – 90 – 90 – 96 – 98

Les quartiles:

Les quartiles sont les valeurs qui divisent une distribution ordonnée en quatre groupes comportant le même nombre de données.

Regarde bien l'exemple suivant.

Trouvons les quartiles de la distribution dont les données sont déjà ordonnées.

5, 7, 7, 8, 10, 12, 12, 13, 16, 18, 20, 21, 21, 22, 23, 24, 27, 29, 30

Il y a trois étapes:

1) Partager la distribution en deux parties égales en trouvant la médiane de la distribution.

Comme nous avons un nombre impair de données, la médiane se trouve :

$$\text{Md} = \frac{(n+1)}{2} = \frac{(19+1)}{2} = 10$$

La médiane est la $10^{\text{ième}}$ donnée et elle vaut 18. Elle représente le deuxième quartile (Q_2).

5, 7, 7, 8, 10, 12, 12, 13, 16, 18, 20, 21, 21, 22, 23, 24, 27, 29, 30

Q_2

2) Maintenant, il faut partager la première moitié en deux parties égales en trouvant la médiane de cette partie. Nous aurons alors les nombres inférieurs à Q_2.

Md = $\frac{(n+1)}{2} = \frac{(9+1)}{2}$ = 5 donc, la cinquième donnée.

La cinquième donnée est 10 et s'appelle le premier quartile (Q_1).

5, 7, 7, 8, $\boxed{10}$, 12, 12, 13, 16, $\boxed{18}$, 20, 21, 21, 22, 23, 24, 27, 29, 30
Q_1 Q_2

3) Partageons la dernière moitié en deux parties égales en trouvant toujours la médiane de cette partie. Nous aurons les nombres supérieurs à Q_2.

Md = $\frac{(n+1)}{2} = \frac{(9+1)}{2}$ = 5 donc, la cinquième donnée.

La cinquième donnée est 23 et s'appelle le troisième quartile (Q_3).

5, 7, 7, 8, $\boxed{10}$, 12, 12, 13, 16, $\boxed{18}$, 20, 21, 21, 22, $\boxed{23}$, 24, 27, 29, 30
Q_1 Q_2 Q_3

Donc, en résumé.

$Q_1 = 10$ $Q_2 = 18$ $Q_3 = 23$

5, 7, 7, 8, $\boxed{10}$, **12, 12, 13, 16**, $\boxed{18}$, **20, 21, 21, 22**, $\boxed{23}$, **24, 27, 29, 30**

1^{er} quart 2^{e} quart 3^{e} quart 4^{e} quart

L'étendue de la distribution est (30 – 5) = 25

L'étendue interquartile est : ($Q_3 – Q_1$) = (23 – 10) = 13

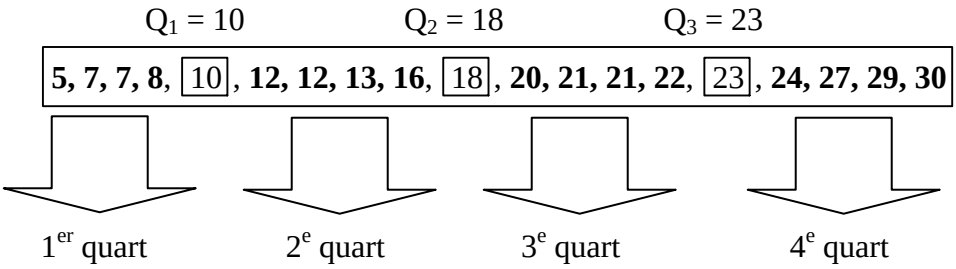
Q_1, Q_2 et Q_3 ne font pas partie des quarts, ils ne font que les séparer.

Chaque quart contient 25% des données.

Le premier quart comprend toutes les données inférieures à Q_1 et le deuxième quart comprend toutes les données situées entre Q_1 et Q_2 et le troisième quart comprend toutes les données situées entre Q_2 et Q_3 et le quatrième quart comprend toutes les données supérieures à Q_3.

Construction d'un diagramme de quartile.

Soit à construire le diagramme de quartile de notre exemple.



Étape 1

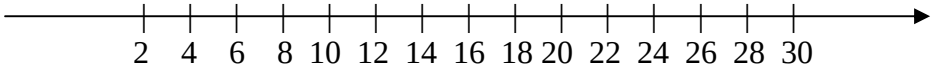
Trouver les quartiles.

Nous avons déjà $Q_1 = 10$, $Q_2 = 18$ et $Q_3 = 23$

Étape 2

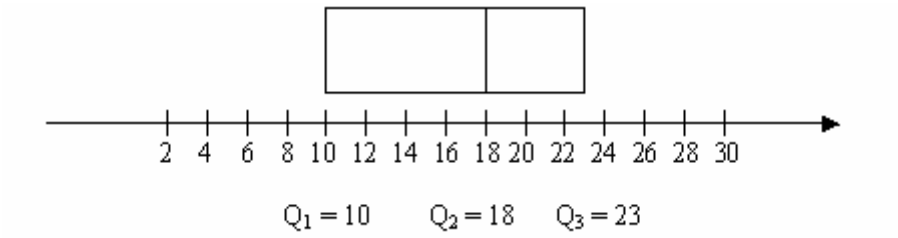
Graduation d'un axe.

Il faut graduer un axe horizontal comprenant la plus petite et la plus grande donnée.



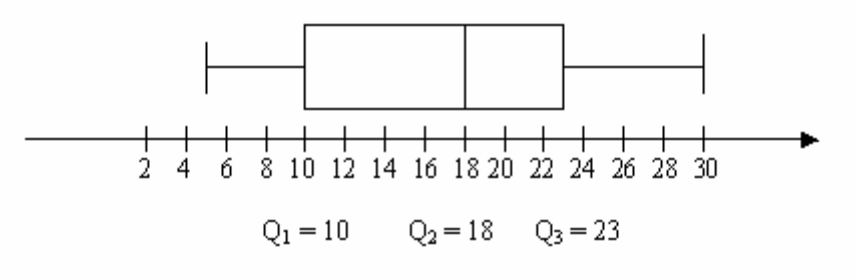
Étape trois

Il faut construire une boite dont les extrémités sont Q_1 et Q_3 et qui est partagée par Q_2.



Étape 4

Finalement, il faut tracer des tiges se rendant aux valeurs minimales et maximales de notre distribution.



Notre diagramme de quartile est maintenant terminé.

Exercice

Voici les résultats d'un examen d'histoire représenté sur un diagramme à feuille.

Notes des élèves à un examen d'histoire

4	7	8							
5	1	4	4	6					
6	0	3	5	7	7	8			
7	0	1	2	3	4	5	8	9	9
8	0	0	4	5	7	8	9		
9	3	5	6	7					
10	0								

<u>Trouve les informations suivantes et demande à ton formateur de vérifier tes réponses.</u>

a) le minimum
b) le maximum
c) l'étendue de la distribution
d) Q_1
e) Q_2
f) Q_3

Rang cinquième

Le rang cinquième correspond au rang du groupe qu'occupe une donnée lorsqu'on divise la population considérée en cinq groupes comportant à peu près le même nombre de données et le premier rang est associé aux données supérieurs (les meilleures).

Exemple:

Voici les résultats d'un mini test de mathématiques qui était noté sur 25 points.

25, 25, 24, 23, 22, 22, 20, 20, 19, 18, 18, 17, 16, 16, 16

Comme nous avons 15 données, nous allons regrouper les données en 5 groupes de 3.

[25, 25, 24]	[23, 22, 22]	[20, 20, 18, 18]	[17, 17, 17]	[16, 16]
1er groupe	2e groupe	3e groupe	4e groupe	5e groupe

Il est à noter que le 3e groupe comporte 4 données. La donnée 18 apparaît deux fois et une donnée ne peut appartenir à deux rangs différents, on a alors placé la deuxième donnée 18 dans le troisième groupe.

L'élève qui a eu une note de 24 se situe dans le 1er rang cinquième et l'élève qui a eu la note de 18 se situe dans le 3e rang cinquième…facile ☺.

Calcul du rang cinquième d'une donnée x quelconque.

Une formule permet de donner le rang cinquième d'une donnée dans une distribution sans toutefois faire les regroupements comme nous l'avons fait dans l'exemple précédent. Cette formule est intéressante lorsque le nombre de données est élevé.

La formule est : $R_5(x) = 5 \times \dfrac{N_{>} + \dfrac{N_{é}}{2}}{N_t}$ où

x représente une donnée de la distribution;
$N_{>}$ représente le nombre de données de valeur supérieur à x;
$N_{é}$ représente le nombre de données d'égale valeur à x;
N_t représente le nombre total de données.

Exemple:

Soit la distribution suivante représentant les derniers résultats d'un examen de mathématiques.

Résultats de l'examen de la première étape en mathématiques

Numéro de fiche	Résultat en %	Numéro de fiche	Résultat en %	Numéro de fiche	Résultat en %
5685	100	1278	85	9442	69
3476	100	4008	84	1648	69
2888	96	3321	79	1690	68
2199	94	7878	77	2778	66
9809	94	9421	76	7222	66
2223	92	0040	76	0003	64
5223	89	3694	73	4567	59
7890	89	2121	71	1145	57
0021	86	8901	71	8899	57
4444	85	5555	69	3456	55

Calculons le rang cinquième de l'élève dont le numéro de fiche est 0003 et qui a obtenu la note de 64%.

La formule est $R_5(x) = 5 \text{ x } \dfrac{N_{>} + \dfrac{N_{é}}{2}}{N_t}$ alors déterminons chacune de ses variables.

x représente la note de notre élève dont la fiche est 0003 soit 64%.
$N_{>}$ représente le nombre de notes supérieures à notre élève soit 25.
$N_{é}$ représente le nombre de notes égales à celle de notre élève soit 1.
N_t représente le nombre total de notes soit 30.

Donc, $x = 64$, $N_{>} = 25$, $N_{é} = 1$ et $N_t = 30$

Alors : $R_5(64) = 5 \text{ x } \dfrac{25 + \dfrac{1}{2}}{30} = 5 \text{ x } \dfrac{25 + 0{,}5}{30} = 5 \text{ x } \dfrac{25.5}{30} = 5 \text{ x } 0{,}85 = 4{,}24.$

Donc, $R_5(64) = 4{,}24$.

Comme le résultat n'est pas entier, nous devons l'arrondir à l'entier supérieur.
Donc, $R_5(64) = 5$.

Notre élève 0003 se situe dans le 5e rang cinquième ☹.

Rang centile

Quand le nombre de données est très grand, il est alors préférable d'utiliser le rang centile plutôt que le rang cinquième. Cela revient à séparer en 100 parties les données d'une distribution.

Le rang centile, nommée R_{100}, indique le pourcentage de données ayant une valeur inférieure ou égale à cette donnée. Les meilleures performances ont un rang centile élevé.

La formule pour calculer le rang centile est similaire à celle utilisée pour calculer le rang cinquième.

La formule est : $R_{100}(x) = 100 \text{ x } \dfrac{N_{<} + \dfrac{N_{é}}{2}}{N_t}$ où

x représente une donnée de la distribution;
$N_<$ représente le nombre de données de valeur inférieure à x;
$N_é$ représente le nombre de données d'égale valeur à x;
N_t représente le nombre total de données.

Comme pour le rang cinquième, nous arrondissons le résultat à l'entier supérieur.

Exemple:

Voici une liste partielle de la masse en kg des 350 élèves de notre école.

$$\underbrace{38, 40, \ldots, 60}_{\text{128 données}}, 61, 64, 64, 65, 67, 70, \underbrace{70, \ldots, 128}_{\text{216 données}}$$

Calculer le rang centile d'un élève qui a une masse de 64 kg.

$R_{100}(x) = 100 \text{ x } \dfrac{N_{<} + \dfrac{N_{é}}{2}}{N_t}$ où $x = 64$, $N_< = 129$, $N_é = 2$, $N_t = 350$

$$R_{100}(64) = 100 \text{ x } \frac{129 + \frac{2}{2}}{350} = 100 \text{ x } \frac{130}{350} = 37{,}15$$

R_{100} = 38% donc 38% des élèves de l'école ont une masse inférieure à 64 kg.

Fin

Ce document n'est qu'une révision *sélective* de la matière contenue dans le cours de MAT- 4104 et ne doit pas être utilisé pour fin de préparation à une évaluation.