
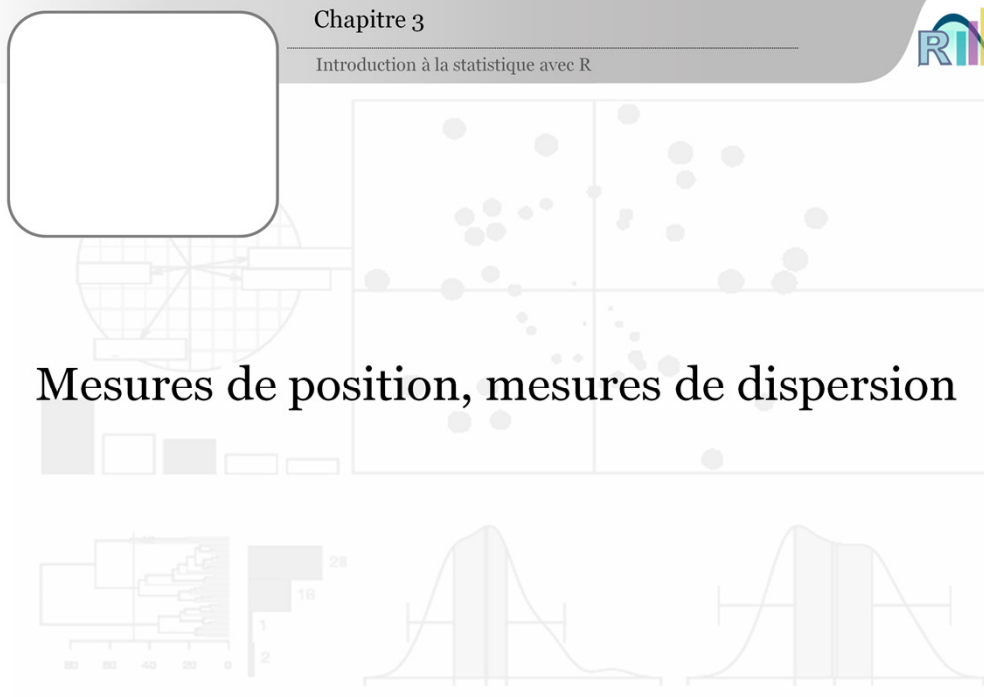



Chapitre 3
Introduction à la statistique avec R






Mesures de position, mesures de dispersion



1


Pr. Bruno Falissard


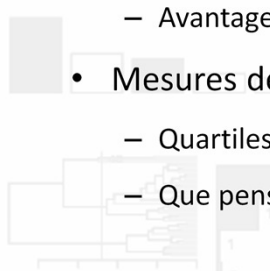
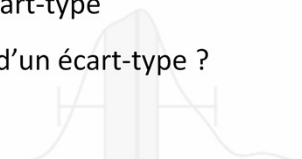



[0:01] Au chapitre précédent, sur les représentations graphiques, nous avons vu qu'un dessin pouvait représenter la distribution d'une variable aléatoire, qu'elle soit qualitative ou quantitative. Le problème c'est que de nos jours, il n'est pas rare qu'un fichier de données compte plusieurs centaines, voire plusieurs milliers de variables aléatoires. Alors on ne peut pas faire 100 ou 1 000 dessins, personne ne va les lire, pas même l'investigateur. On a donc besoin de mesures agrégées qui synthétisent l'information, qui la résument. De là, la notion de mesures de position et de mesures de dispersion, parmi lesquelles les très connues moyennes et écarts-types.


Plan


Introduction à la statistique avec R > Mesures de position, de dispersion



- Mesures de position
 - Moyenne, médiane
 - Avantages et inconvénients
- Mesures de dispersion
 - Quartiles, écart-type
 - Que penser d'un écart-type ?


2

Pr. Bruno Falissard


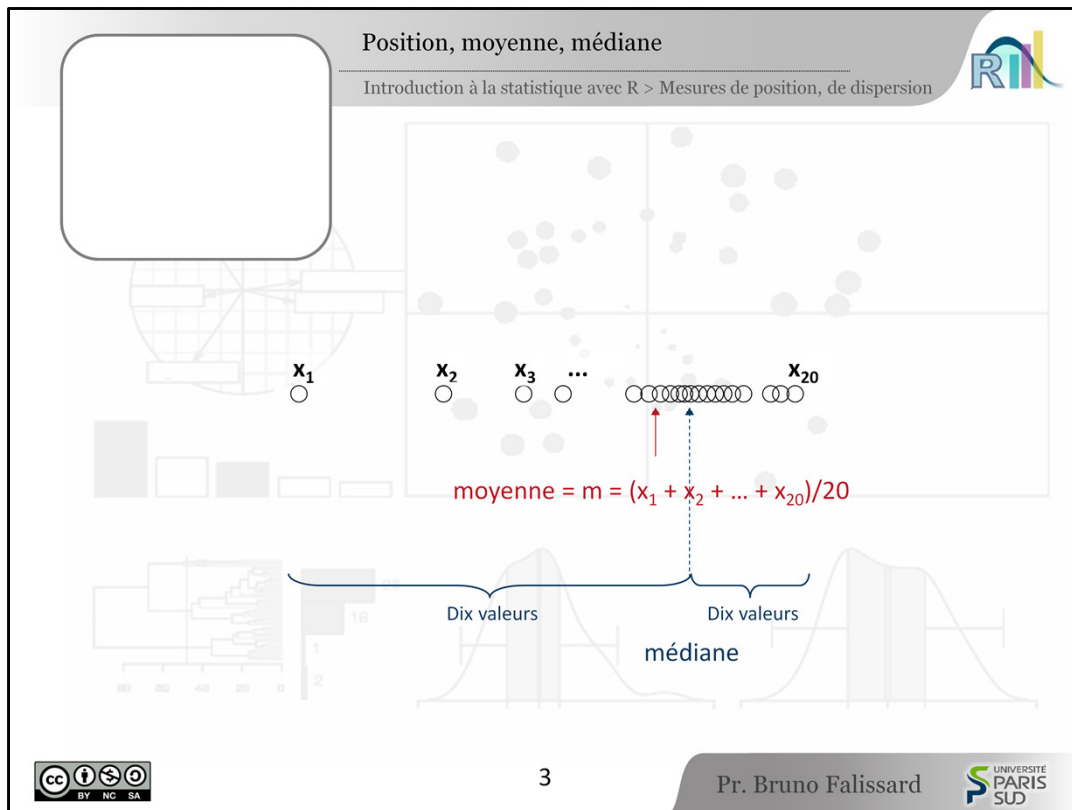
[0:41] Imaginez que vous ayez à travailler sur une étude visant à évaluer la consommation de cannabis en France. Selon que la population à étudier a un âge qui se situe autour de 15 ans ou un âge autour de 60 ans, la problématique de santé publique sous-jacente, les déterminants de consommation vont être complètement différents. Il est donc capital de pouvoir disposer d'un paramètre simple qui permet de dire immédiatement autour de quel âge se situe la population. On appelle ça une **mesure de position**.

En complément de cette mesure de position, on a besoin de connaître la **dispersion**. Si vous prenez des adolescents ; si vous les interrogez en classe de 3^{ème}, ils auront tous autour de 15 ans, ils seront très homogènes ; si vous les interrogez en mélangeant collège et lycée, vous aurez des jeunes dont l'âge varie entre 11 et 18 ans, la problématique sera complètement différente. Nous allons donc voir dans ce chapitre des mesures de position et des mesures de dispersion.

Pour le cas de variables catégorielles, une mesure de position, c'est très simple. Il suffit de lister le pourcentage de toutes les modalités, de toutes les catégories de la variable qualitative ou catégorielle évaluée.

Pour une variable quantitative, c'est un tout petit peu plus compliqué et vous savez qu'il existe deux paramètres : la **moyenne** et la **médiane**. Nous allons les voir en détail.

En ce qui concerne les mesures de dispersion, les **quartiles** sont utilisés ainsi que l'**écart-type** très connu mais pas forcément à bon escient. Vous allez voir, son interprétation est délicate.



[2:14] Considérons donc une variable avec une vingtaine d'observations. L'objectif est de trouver une valeur autour de laquelle se situe globalement l'ensemble des observations.

La première solution, la plus classique, est de calculer la **moyenne**, c'est-à-dire de faire la somme des observations et de diviser par 20. On est conditionné depuis l'école primaire à avoir la moyenne de nos résultats. Au baccalauréat, il y a la moyenne des notes, une moyenne pondérée d'ailleurs. Bref, dans la vie de tous les jours on utilise souvent des moyennes et on ne se rend même plus compte qu'en réalité le sens à donner au mot "moyenne" n'est pas si évident que ça. Si je vous dis que vous avez eu 15 notes et que la moyenne de ces 15 notes c'est 12, qu'est-ce que ça signifie ? Ce n'est pas si évident que ça. Vous ne savez pas par exemple, s'il y a plus de notes supérieures à 12 que de notes inférieures à 12, ou le contraire. Une façon de se représenter géométriquement ce que c'est qu'une moyenne, c'est de considérer qu'elle est le centre de gravité, le barycentre de l'ensemble des observations. La moyenne a ainsi un sens physique. Elle n'est pas le seul paramètre que l'on peut utiliser.

Il existe aussi la **médiane**. La médiane a un sens beaucoup plus clair. 50% des observations lui sont plus petites, 50% des observations lui sont plus grandes. Si la distribution est symétrique, alors la médiane est égale à la moyenne. Dans le cas contraire, et c'est la situation présente, où on voit que les observations sont beaucoup plus dispersées à gauche qu'à droite, alors il y a des différences non négligeables entre la moyenne et la médiane.

Moyenne *versus* Médiane

Introduction à la statistique avec R > Mesures de position, de dispersion

- Médiane :
 - Intuitif
 - Robuste
- Moyenne :
 - Simple à calculer
 - Barycentre
 - Propriété « comptable »

4

Pr. Bruno Falissard


[3:53] Alors, en pratique faut-il utiliser des médianes ou des moyennes ? Essayons de voir le pour et le contre.

En faveur de la médiane, il y a d'abord son caractère intuitif. Au moins, on sait ce que ça veut dire. 50% des valeurs sont plus petites que la médiane, 50% des observations sont plus grandes. Il y a aussi la robustesse, la médiane est très peu sensible aux valeurs extrêmes. Imaginez par exemple que vous ayez un jeu de données avec 50 sujets. Vous prenez l'âge de ces sujets, et puis il y a une erreur de saisie, au lieu que ce soit 25 ans, il y a 2500 ans. Ça ne va rien changer à la médiane. Puisque la médiane coupe l'échantillon en deux, que les valeurs extrêmes soient hyper-extrêmes, ou un petit peu extrêmes ça ne changera rien à la médiane alors que naturellement ça ne sera pas du tout la même chose pour la moyenne. La médiane est donc un paramètre robuste.



Alors **en faveur de la moyenne**, d'abord, elle est simple à calculer. On fait la somme, on divise par "n", alors qu'une médiane ce n'est pas du tout pareil. En gros pour calculer une médiane, il faut classer quasiment toutes les observations. Et classer à la main 1000 observations, franchement c'est l'enfer. Bien sûr aujourd'hui, il y a les ordinateurs, mais il faut comprendre que quand ont été développés ces paramètres, il n'existait pas d'ordinateur. Alors qui plus est, la moyenne elle a une propriété géométrique et physique, c'est un barycentre, c'est un centre de gravité, et autant la médiane n'est pas un beau paramètre en termes mathématiques, la moyenne a des propriétés mathématiques intéressantes. Au-delà de ses propriétés mathématiques, la moyenne a un certain intérêt comptable.


Moyenne *versus* Médiane


Introduction à la statistique avec R > Mesures de position, de dispersion



- Médiane :
 - Intuitif
 - Robuste
- Moyenne :
 - Simple à calculer
 - Barycentre
 - Propriété « comptable »

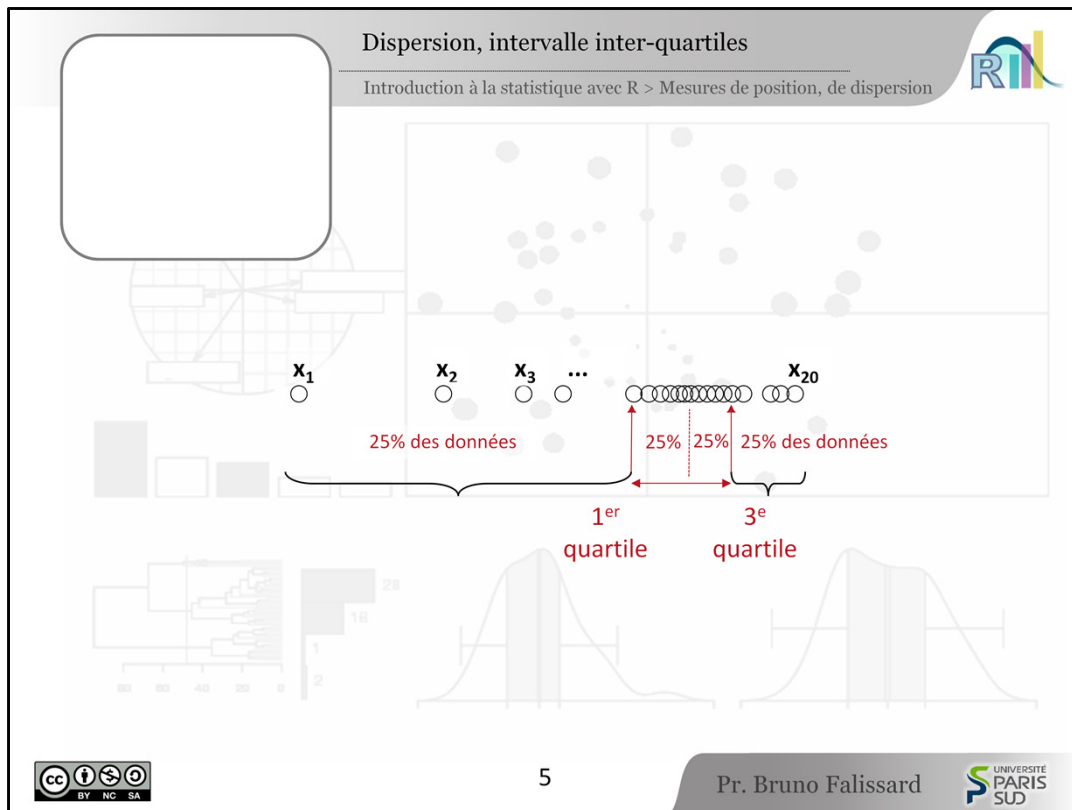




4'

Pr. Bruno Falissard


[5:27] On dit traditionnellement que pour représenter une variable qui a une distribution asymétrique, assez fortement asymétrique, il faudrait utiliser une médiane et pas une moyenne ; et bien pas toujours.

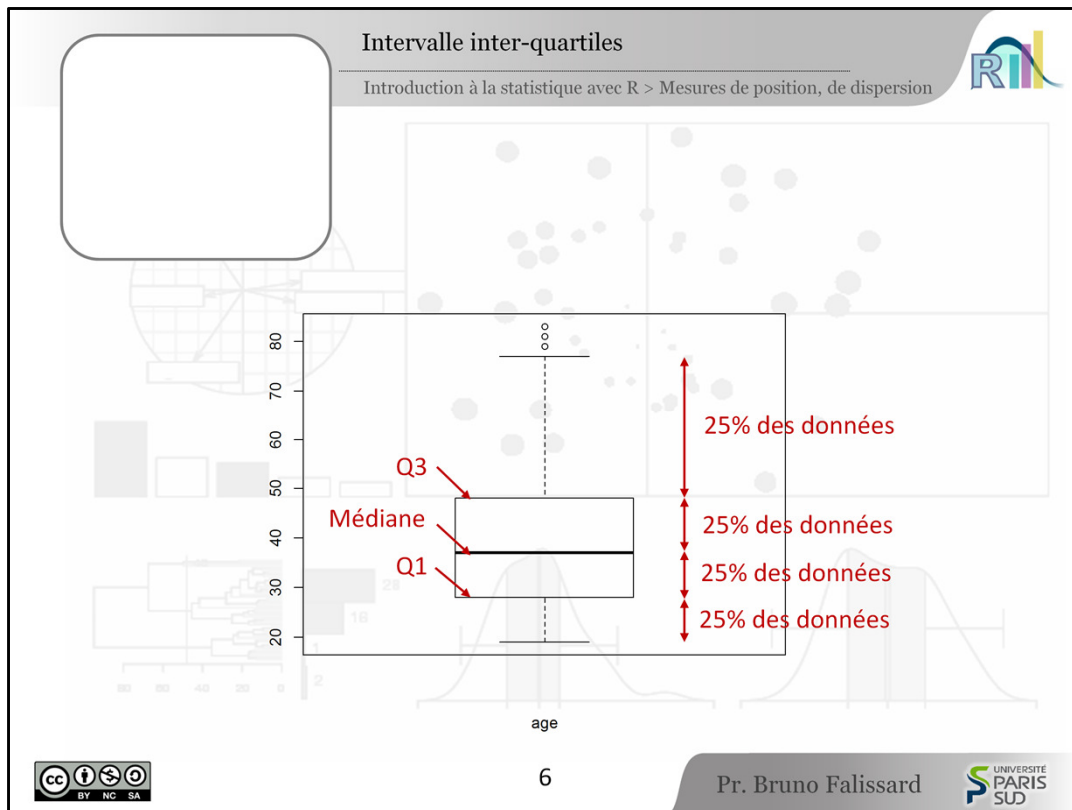
Prenons un exemple médical, la distribution de la durée de séjour de patients hospitalisés. Les patients en général sont hospitalisés quelques jours. Et puis quelques-uns d'entre eux vont rester, à cause de complications, plusieurs semaines, voire plusieurs mois, mais ils sont très rares. Il serait logique donc, pour essayer d'avoir un paramètre de position de la durée de séjour d'avoir la médiane. Hors les directeurs d'hôpitaux s'intéressent à la durée moyenne de séjour. Pourquoi ? Parce que si vous avez la durée moyenne de séjour, et que vous avez le nombre de séjours hospitaliers dans l'année, quand vous multipliez l'un par l'autre, vous avez le nombre de jours de lits d'hospitalisation occupés dans l'année. Et si l'hôpital est payé au nombre de jours de lits d'hospitalisation occupés, alors le directeur d'hôpital sait exactement combien il va gagner pour son année.



[6:27] Voyons maintenant les paramètres de dispersion. Nous retrouvons nos 20 observations, et dans un premier temps nous nous penchons sur l'**intervalle interquartiles**.

La médiane découpait le jeu d'observations en deux parts égales. Les quartiles découpent le même jeu d'observations en quatre parts égales.

- Il y a 25% des sujets qui ont des valeurs inférieures au 1^{er} quartile,
- 25% des sujets qui ont des observations supérieures au 3^{ème} quartile,
- et donc entre le 1^{er} et le 3^{ème} quartile, nous avons 50% des observations, ce sont les 50% qui se regroupent autour de la médiane. L'intervalle interquartile a ainsi un sens clair et immédiat, il regroupe la moitié de l'échantillon qui se situe autour de la médiane, et d'ailleurs, c'est pour cette raison que dans le boxplot...




[7:22] ... dans la boîte à moustaches que nous avons vue dans le chapitre précédent, la boîte de la boîte à moustaches était constituée

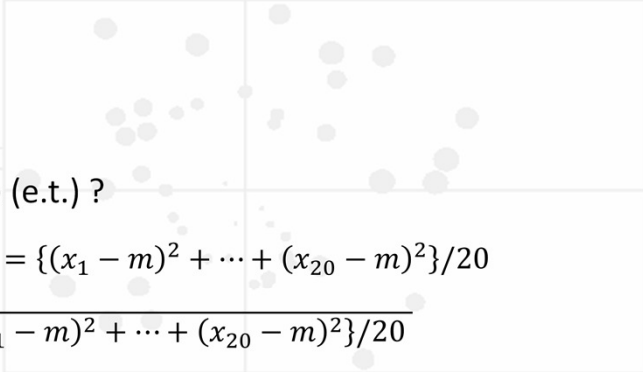
- dans sa borne inférieure du 1^{er} quartile*,
- au milieu de la médiane
- et dans sa borne supérieure du 3^{ème} quartile.

* Erreur dans la vidéo


Ecart-type

Introduction à la statistique avec R > Mesures de position, de dispersion






- Et l'écart-type (e.t.) ?
 - $e.t.^2 = Var = \{(x_1 - m)^2 + \dots + (x_{20} - m)^2\}/20$
 - $e.t. = \sqrt{\{(x_1 - m)^2 + \dots + (x_{20} - m)^2\}/20}$
- Pourquoi l'écart-type ?
 - Une inertie
 - $$e.t.^2 = Var = \frac{(x_1^2 + \dots + x_{20}^2)}{20} - \left(\frac{x_1 + \dots + x_{20}}{20}\right)^2$$



7

Pr. Bruno Falissard



[7:39] L'intervalle interquartiles n'est pas le paramètre de dispersion le plus utilisé. Le champion, c'est l'écart-type, vous êtes tous au courant. Alors, quelle est la définition d'un écart-type ?

L'**écart-type**, par définition, c'est la racine carrée de la variance, et la **variance**, c'est la moyenne des carrés des écarts à la moyenne. Et donc l'écart-type, c'est la racine carrée de la moyenne des carrés des écarts à la moyenne. Qu'est-ce que ça veut dire en pratique ?

Ça ne veut rien dire ! L'écart-type n'a aucun sens intuitif. Alors pourquoi utilise-t-on un tel paramètre ? Une première explication c'est que, comme la moyenne, l'écart-type a une certaine interprétation physique. Il correspond à une inertie. De la même façon, l'écart-type a d'excellentes propriétés mathématiques, d'ailleurs on le retrouve dans le `test t` que nous verrons dans un cours suivant, et c'est complètement différent de l'intervalle interquartiles, qui lui est comme la médiane, n'a pas du tout de bonnes propriétés mathématiques. Enfin, l'écart-type a une propriété qui le rend très utile et très facile à calculer, l'écart-type au carré, c'est-à-dire la variance, c'est la moyenne des carrés moins la moyenne au carré, et nous allons voir qu'au début du XX^{ème} siècle, cette propriété était fantastique.

Ecart-type

Introduction à la statistique avec R > Mesures de position, de dispersion

• Comment calculer une variance ?

x_i	Σx_i	x_i^2	Σx_i^2
x_1	s_1	x_1^2	q_1
x_2	s_2	x_2^2	q_2
...			
x_{1000}	s_{1000}	x_{1000}^2	q_{1000}

$$\text{Var} = q_{1000} / 1000 - (s_{1000} / 1000)^2$$

8

Pr. Bruno Falissard

[9:02] Imaginez donc que vous êtes un statisticien de l'armée et que vous travaillez sur la taille, le poids, etc. des conscrits, c'est-à-dire des jeunes gens qui viennent faire leur service militaire. Votre travail en particulier est de calculer la moyenne et l'écart-type des conscrits et vous pouvez avoir un échantillon de grande taille, notamment de plusieurs milliers de sujets. Alors comment faisait-on à l'époque pour calculer la moyenne et l'écart-type d'un échantillon constitué de 1000 individus ? On dressait un tableau de 4 colonnes et 1000 lignes.

- sur la 1^{ère} colonne, toutes les tailles,
- sur la 2^{ème} colonne, les sommes successives des tailles, c'est-à-dire $s_2 = x_2 + x_1$, et $s_3 = x_3 + s_2$ (qui est égal à $x_2 + x_1$) donc $s_3 = x_3 + x_2 + x_1$, etc. jusqu'à $s_{1000} = x_1 + \dots + x_{1000}$, s_{1000} vaut bien la somme de toutes les tailles,
- la 3^{ème} colonne, c'est les tailles au carré,
- la 4^{ème} colonne, c'est les sommes cumulées des tailles au carré.

On obtient ainsi q_{1000} et s_{1000} et grâce à la formule de la variance, et après l'extraction d'une racine carrée, qui n'était pas simple à l'époque, mais là, il n'y en a qu'une à faire, il n'y en a pas 1000, on pouvait calculer l'écart-type.

Alors imaginons maintenant que, en fin de journée vous vous rendez compte qu'il n'y avait pas mille individus, il y en avait 1001.

Ecart-type

Introduction à la statistique avec R > Mesures de position, de dispersion

- Comment calculer une variance ?

x_i	Σx_i	x_i^2	Σx_i^2
x_1	s_1	x_1^2	q_1
x_2	s_2	x_2^2	q_2
...			
x_{1000}	s_{1000}	x_{1000}^2	q_{1000}
x_{1001}	s_{1001}	x_{1001}^2	q_{1001}

$$\text{Var} = q_{1000} / 1000 - (s_{1000} / 1000)^2$$

$$\text{Var} = (q_{1000} + x_{1001}^2) / 1001 - \{(s_{1000} + x_{1001}) / 1001\}^2$$


8'

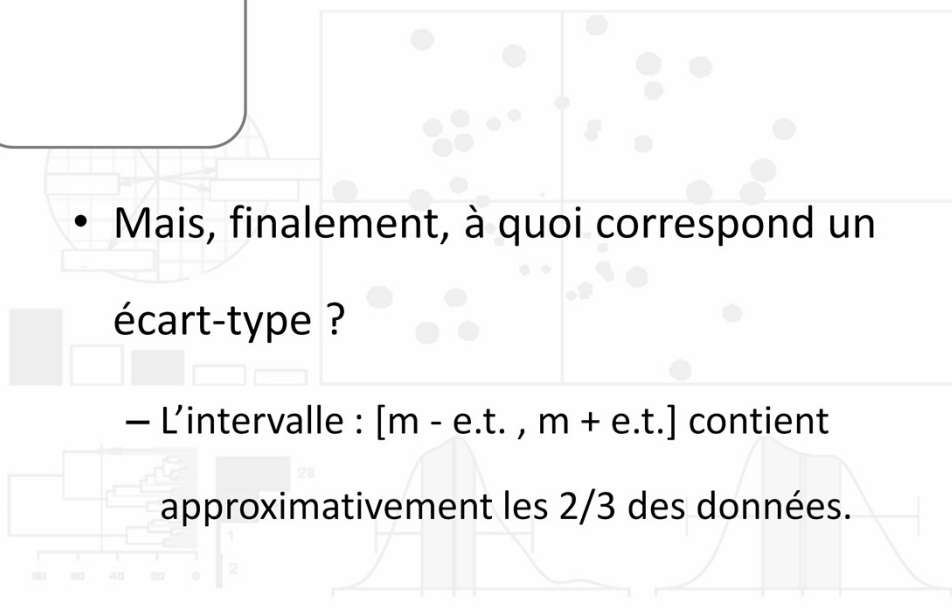
Pr. Bruno Falissard

[10:28] Est-ce que c'est une grosse catastrophe ou est-ce que ce n'est pas un problème? Si la définition de l'écart-type avait été différente, par exemple, une définition intuitive, on aurait pu imaginer que l'écart-type ça aurait été la moyenne des valeurs absolues des écarts à la moyenne ; ainsi l'écart-type ça aurait été vraiment l'écart moyen par rapport à la moyenne, ça a un sens. Mais si l'écart-type avait été égal à ce paramètre, alors le fait de rajouter une 1001^{ème} observation vous obligerait de refaire au moins 1000 calculs, ce qui était absolument inconcevable. Alors qu'avec la définition que l'on a de la variance, c'est-à-dire la moyenne des carrés moins la moyenne au carré, si vous rajoutez un 1001^{ème} individu, il suffit de rajouter x_{1001}^2 à q_{1000} , de rajouter x_{1001} à s_{1000} donc deux additions, vous divisez par 1001, vous avez la variance, une racine carrée supplémentaire, et l'écart-type. Vous imaginez donc le succès qu'a eu l'écart-type quand il fallait faire tous les calculs à la main.


Ecart-type

Introduction à la statistique avec R > Mesures de position, de dispersion






- Mais, finalement, à quoi correspond un écart-type ?
- L'intervalle : $[m - e.t. , m + e.t.]$ contient approximativement les 2/3 des données.



9

Pr. Bruno Falissard
 

[11:31] Alors au total, l'écart-type est principalement utilisé pour ses propriétés mathématiques et calculatoires et pas par le sens qu'on peut lui donner. Est-ce que c'est grave ?

Pas vraiment parce qu'il y a une possibilité de se rattraper. Si la variable a une distribution normale, l'intervalle qui contient la moyenne moins l'écart-type jusqu'à la moyenne plus l'écart-type représente approximativement les 2/3 des données. Et c'est ça que les gens ont en tête, c'est que $\text{moyenne} \pm \text{écart-type}$ ça représente le gros de la troupe des données.

Et en conclusion, si formellement un écart-type ça ne veut rien dire, en réalité la plupart des utilisateurs comprennent bien le sens à donner à ce paramètre.