université PARIS DIDEROT

Méthodes statistiques TD n°7-8

O. Dadoun dadoun@in2p3.fr semaine du 27/10/2015 - 10/11/2015

Exercice

But: introduire une nouvelle caractéristique d'une série statistique pour mesurer sa dispersion autour de la moyenne.

Soit le relevé des températures minimales et maximales d'une région:

Date	lundi	mardi	mercredi	jeudi	vendredi	samedi	dimanche
T° (min)	8.8	12.2	13.5	12.7	8.5	7.7	5.2
T° (max)	19.5	19.9	18.6	17.8	18	17.3	18.1

1) Calculer la température minimale moyenne $\overline{x} = 9.8$
2) Compléter la différence à la moyenne. Qu'observe-t-on quand on calcule la moyenne de ces différences ?

T° (min)	8.8	12.2	13.5	12.7	8.5	7.7	5.2
Écart $x_i - \overline{x}$	-1.0	2.4	3.7	2.9	-1.3	-2.1	-4.6

3) De 2) on amène à considérer que des quantités positives, pour cela on utilise le carré de l'écart à la moyenne. Compléter le tableau et calculer la moyenne des carrées des écarts à la moyenne 56.12/7= 8.01
4) Le nombre obtenu s'appelle la variance. Pour compenser l'utilisation du carré on calcule la racine carrée de la variance qu'on appelle l'écart type 2.83.

T° (min)	8.8	12.2	13.5	12.7	8.5	7.7	5.2
Écart $x_i - \overline{x}$	-1.0	2.4	3.7	2.9	-1.3	-2.1	-4.6
Carré de l'écart $(x_i - \overline{x})^2$	1.00	5.76	13.69	8.41	1.69	4.41	21.16

Si l'écart type donne une bonne indication sur la dispersion de la série il n'est malheureusement par interprétable ou représentable aussi facilement que les quartiles.
L'écart-type représente la distance moyenne des observations par rapport à la moyenne

Quelques formules

- Moyenne $\overline{x} = \frac{\sum x}{N}$
- Variance $s^2 = \frac{\sum (x-\overline{x})^2}{N} = \frac{\sum x^2}{N} - \overline{x}^2 = \overline{x^2} - \overline{x}^2$
- Écart-type s = $\sqrt{variance}$

N.B: Plus l'écart-type est faible, plus les données sont dispersées autour de la moyenne. Inversement, plus l'écart-type est élevée, plus les données sont dispersées loin de la moyenne.

Définition:
On note N l'effectif total d'un série, x_i les valeurs du caractère et n_i les effectifs correspondants.
La moyenne s'écrit $\bar{x} = \sum_{i=1}^{i=p} \frac{n_i x_i}{N}$ et la variance s'écrit $\sum_{i=1}^{i=p} \frac{n_i (x_i - \bar{x})^2}{N} = \sum_{i=1}^{i=p} \frac{n_i x_i^2}{N} - \bar{x}^2$

- Coefficient de la variation
  $C_v = \frac{S}{\bar{x}}$ (en pourcentage)

Le coefficient de variation permet de comparer la variabilité de deux séries qui ont des moyennes très différentes. Comme ce dernier s'exprime en pourcentage (%) son utilisation est plus simple que celui de l'écart type.

Grosso modo
Entre 0 et 15%: on dira que la moyenne est représentative de la distribution;
Entre 15% et 30%: on dira qu'il faut utiliser avec prudence la moyenne de la distribution;
À 30% et plus : On dira que la moyenne n'est pas représentative de la distribution. C'est à dire qu'il y a des observations qui se situent très loin de la moyenne.

Exercice
On a demandé à 50 personnes prenant l'autobus, le nombre de fois où chacune de ces personnes a utilisé ce type de transport pendant la semaine écoulée.

Nb de voyages en bus x_i	1	2	3	4	5	6	7	8	9	10
Effectif n_i	3	3	5	7	6	9	5	4	5	3
$n_i * x_i$	3	6	15	28	30	54	35	32	45	30
Carré de l'écart $n_i * (x_i - \bar{x})^2$/N	1.24	0.76	0.65	0.34	0.03	0.03	0.20	0.47	1.18	1.18
ou $n_i x_i^2$	9	12	45	112	150	324	245	256	405	300

Déterminer la valeur moyenne du nombre de voyage en bus 5.56
Compléter la dernière ligne du tableau.
Calculer ensuite l'écart-type.
Attention: les effectifs ne sont pas tous égaux à 1 !

$\sum_{i=1}^{i=p} \frac{n_i (x_i - \bar{x})^2}{N} = 6.12$ => s=2.47
ou
$\sum_{i=1}^{i=p} \frac{n_i x_i^2}{N} - \bar{x}^2$ = 37.04 - 5.56^2
Le nombre de trajet moyen est donc 5.56 +/- 2.47

Exercice
On présente au sujet un écran sur lequel sont projetés 5 nombres, puis un second écran avec un seul nombre. Le sujet doit appuyer sur un bouton réponse positive (+) s'il pense que le nombre présenté sur le second écran faisait partie des nombres présentés sur le premier écran et sur un bouton réponse négative (–) dans le cas contraire.

Complétez ci-après le tableau des temps de réaction (en centièmes de seconde) de 100 sujets à l'une des épreuves de reconnaissance de stimulus, intervenant dans une étude sur la mémoire à court terme.

	Temps	x_i	n_i	$n_i x_i$	$n_i x_i^2$
1	[40-50[	45	4	180	8100
2	[50-60[	55	19	1045	57475
3	[60-70[	65	41	2665	173225
4	[70-80[	75	21	1575	118125
5	[80-90[	85	11	935	79475
6	[90-100[	95	3	285	27075
7	[100-110[	105	1	105	11025
$\sum$			100	6790	474500

Calculer la variance, écart type et coefficient de variation.
Quelles informations ces paramètres vous fournissent-ils sur la dispersion des temps de réaction de cette distribution ?

$\bar{x} = \frac{\sum n_i x_i}{N} = 67.90$

$\frac{\sum n_i x_i^2}{N} = 474500/100$

$-s^2 = 4745 - 67.90^2 = 134.59$ centièmes de secondes au carré

- s = 11.60 centièmes de secondes
(- Cv = 17.1%)

Variance : aucune signification concrète.
Écart type : les temps de réaction de l'ensemble des sujets se repartissent avec une distance moyenne à la moyenne de 11,60 centièmes de seconde (on peut approcher la notion en disant que ces temps se repartissent avec une moyenne de 69,7 centièmes de seconde ± 11,60 centièmes de seconde).
Coefficient de variation: l'écart type représente 17,1% de la valeur moyenne (dispersion moyenne pour une épreuve standardisée).

Exercice
Série à caractère continu.
Soit le montant des achats de fournitures mensuel d'une entreprise

Montant des achats (€)	[0 ; 20[	[20 ; 40[	[40 ; 60 [	[60 ; 100[	[100 ; 140 [	[140 ; 200[
x_i	10	30	50	80	120	170
Effectifs	35	41	30	12	5	2

$n_i x_i^2$	3500	36900	75000	76800	72000	57800

N=125

En utilisant les centres des classes calculer moyenne puis la variance.

$\overline{x} = \sum_{i=1}^{i=p} \frac{n_i x_i}{N}$ = 39.84

$s^2 = \sum_{i=1}^{i=p} \frac{n_i x_i^2}{N} - \overline{x}^2$ = 2576. - 1587.22 => s=31.44

<u>Exercice</u>

Considérons une entreprise E comportant 2 établissements: E1 et E2 qui emploient chacun 200 salariés. Au sein de E1 le salaire moyen est de 1500 euros avec un écart type de 800. Au sein de E2 le salaire moyen est de 2500 euros avec un écart type de 1100. Dans quelle entreprise le salaire est-il le plus dispersé ?

Comme les 2 salaires moyens sont différents on peu pas comparer directement les écart-types. Afin de comparer les 2 indicateurs de dispersions autour de la moyenne il faut comparer la dispersion relative c.a.d le coefficient de variation.

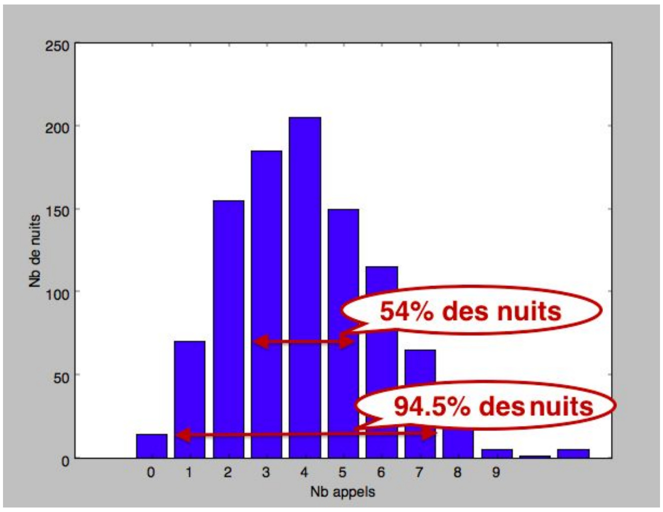
CV1=800/1500=53,3% et CV2= 44% , les salaires sont donc davantage dispersés dans l'établissement E1.

<u>Exercice</u>

Un hôpital de garde a enregistré le nombre d'appels reçus pendant pendant 1000 nuits entre 1h et 2h du matin.

Nb appels x_i	0	1	2	3	4	5	6	7	8	9	10	11
Nb du nuits n_i	14	70	155	185	205	150	115	65	30	5	1	5
$n_i x_i$	0	70	310	555	820	750	690	455	240	45	10	55
$n_i x_i^2$	0	70	620	1665	3280	3750	4140	3185	1920	405	10	605

- Déterminer la moyenne et l'écart type de cette série. $\overline{x} = 4$ $s^2 = \sum_{i=1}^{i=p} \frac{n_i x_i^2}{N} - \overline{x}^2$ = 19.65-16=3.65 => s= 1.91
- Déterminer le nb de nuits pour lesquelles le nb d'appels appartient à l'intervalle $[\overline{x} - s; \overline{x} + s]$; quelle est la fréquence correspondante ?
  [2.08;5.91] c.a.d x_i >2 & x<6 185+205+150= 540 nuits soit 54%
- Même question avec l'intervalle $[\overline{x} - 2s; \overline{x} + 2s]$.
  [0.18;7.82] c.a.d x_i >0 & x<8 70+155+185+205+150+115+65= 945 nuits soit 94.5%
  N.B. : cet intervalle dont l'amplitude est de 4 fois l'écart type contient presque la totalité des effectifs.
- Faites l'histogramme où figura les 2 précédents intervalles



Lorsque le caractère statistique a une distribution normale gaussienne, grossièrement en forme de cloche, l'écart type prend tout son sens.

- Dans l'intervalle [$\overline{x} - s; \overline{x} + s$], on trouve 68 % de la population.
- Dans l'intervalle [$\overline{x} - 2s; \overline{x} + 2s$], on trouve 95 % de la population.
- Dans l'intervalle [$\overline{x} - 3s; \overline{x} + 3s$], on trouve 99,7 % de la population.