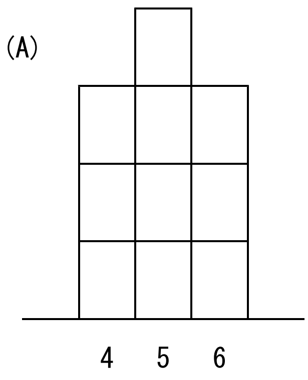
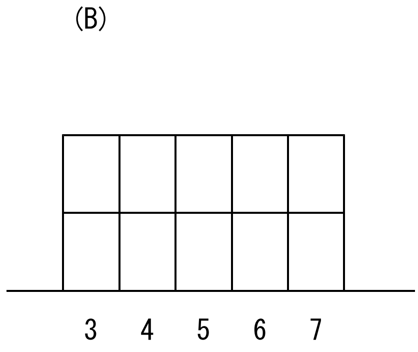
資料4-3

【分散(variance)と標準偏差(SD:standard deviation)のわかりやすい説明】

●下の２つの分布はどちらも人数が10人(N=10)，平均値と中央値(Md)はいずれも５である。しかし，広がり具合が違い，Ｂの方がＡよりも広がっている。この広がり(散らばり)の違いをどのように表現(数量化)したら良いだろうか？





(1)そこで，一人一人が平均(=5)からどの位離れているかを求めてみる。

(A)	(B)
4-5=-1	3-5=-2
4-5=-1	3-5=-2
4-5=-1	4-5=-1
5-5= 0	4-5=-1
5-5= 0	5-5= 0
5-5= 0	5-5= 0
5-5= 0	6-5= 1
6-5= 1	6-5= 1
6-5= 1	7-5= 2
+)　6-5= 1	+)　7-5= 2
0	0

それぞれの全体的大きさを比較するために合計すると０になってしまう(平均値の定義から当然そうなる)。

(2)そこで，各値を２乗してから合計してみると

(A)	(B)
$(4-5)^2=(-1)^2$	$(3-5)^2=(-2)^2$
$(4-5)^2=(-1)^2$	$(3-5)^2=(-2)^2$
…	…
…	…
+)　$(6-5)^2=1^2$	+)　$(7-5)^2=2^2$
6	20

となり，(B)の方が大きくなる。しかし，この場合はdata数が同じ(10個)だから比較できるが，data数が異なると大きさの比較ができない。そこで，これらの値をdata数で割る(この場合，data数－１の値で割ったものを使う：不偏分散)。

6/(10-1)=2/3　　　　　　20/(10-1)=20/9

これらの値を分散(variance)という。そして，この値の平方根を標準偏差(SD:standard deviation)という。

●チェビシェフの不等式

標準偏差(SD)の値には，次のような性質がある(x_mean は平均，s_xはSDを表す)。

$$P\{\,|x_i - x\_mean| \leqq k \cdot s_x\,\} > 1-1/k^2$$

具体的には，ｋ＝２　とすると，平均から±２s_x の中に，全dataの75%以上が含まれることになる。この性質は分布の形とは関係せずに成立する。