1.3　分散と標準偏差

分散とは、標準偏差とは？

まずは復習から。

n 個の**量的変数**の集合であるデータ

$x_1,\ x_2,\ x_3,\ \cdots,\ x_n$

に対して，次の計算で得られる**代表値**を，**平均値**といいます。

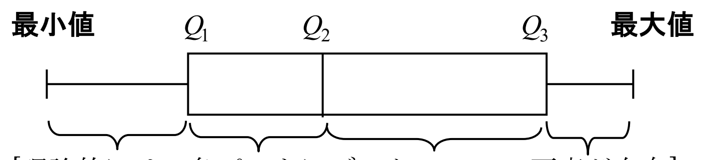
［平均値］　$\mu = \dfrac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} x_k$

データの各要素と平均値[又は代表値]との差を，**偏差**といいます。

［偏　差］　$x_k - \mu$　　（$k = 1,\ 2,\ 3,\ \cdots,\ n$）

※偏差は，データの要素の散らばり具合を示す用語です。
　前回の箱ひげ図では，**四分位偏差**を学習しましたね。

【箱ひげ図の場合：データの要素の散らばり具合】



[理論的には，各パートにデータの 25%の要素が存在]

但し　Q_1：**第１四分位**
　　　Q_2：**第２四分位**(中央値)
　　　Q_3：**第３四分位**

偏差の総和は，0 になるため(次頁の【研究】参照)

通常は，「偏差の 2 乗」の総和を，データ数 n で割った値を用います。

この計算のことを，**分散**といい，記号 σ^2（読み：シグマ　２乗）で表す。

（※ギリシア文字「シグマ」　小文字　σ，　大文字　$\sum$　）

［分散］　$\sigma^2 = \dfrac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + (x_3 - \mu)^2 + \cdots + (x_n - \mu)^2}{n} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(x_k - \mu)^2$

分散の平方根をとったものを，**標準偏差**と言い，記号σで表す。

［標準偏差］　**標準偏差** $=\sqrt{\textbf{分散}}$

※標準偏差(又は分散)の値が小さいほど，データのばらつきが少ないと言われます。

例題　次の10個のデータの平均値，偏差の総和，分散と標準偏差を求めよ。

32　33　28　27　34　22　37　23　34　30

［解答］　平 均 値　$\mu=\dfrac{32+33+28+27+34+22+37+23+34+30}{10}=\dfrac{300}{10}=30$

偏差総和　$(32-30)+(33-30)+(28-30)+(27-30)+(34-30)$
$+(22-30)+(37-30)+(23-30)+(34-30)+(30-30)$

$=\boxed{2+3+(-2)+(-3)+4+(-8)+7+(-7)+4+0}=0$

偏差の2乗の総和をとる

分　散　$\sigma^2=\dfrac{\boxed{2^2+3^2+(-2)^2+(-3)^2+4^2+(-8)^2+7^2+(-7)^2+4^2+0^2}}{10}$

$=\dfrac{4+9+4+9+16+64+49+49+16+0}{10}=\dfrac{220}{10}=22$

標準偏差　$\sigma=\sqrt{22}(=4.7)$

問3.5　次のデータの平均値，分散と標準偏差を求めよ。

(1) 6　4　8　5　2　　　(2) 4　7　1　9　5　0　6　8

＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝

【研究：偏差の和は０】

$$\sum_{k=1}^{n}(x_k-\mu)=(x_1-\mu)+(x_2-\mu)+(x_3-\mu)+\cdots+(x_n-\mu)$$

$$=(x_1+x_2+x_3+\cdots+x_n)-n\ \mu$$

$$=n\left(\boxed{\frac{x_1+x_2+x_3+\cdots+x_n}{n}}-\mu\right)$$

[※平均の計算式]

$$=n(\mu-\mu)=0$$

＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝

1.4　チェビシェフの不等式

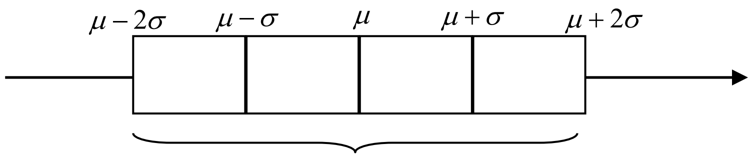
チェビシェフの不等式とは？

$\mu - k\sigma$ から $\mu + k\sigma$ には，データの要素の少なくとも

［チェビシェフの不等式］　$100 \times \left(1 - \dfrac{1}{k^2}\right)$　[%]

が存在する。

例）$k = 2$ のとき　$100 \times \left(1 - \dfrac{1}{k^2}\right) = 100 \times \left(1 - \dfrac{1}{4}\right) = \dfrac{300}{4} = 75\%$



［※データの全要素の数なくとも 75%が存在］

1.5　度数分布表

度数分布とは？

○ある変量のデータを分類したリストを**度数分布**という。このとき

1）各小区間を**階級**という。

2）階級の中央の値を**階級値**という。

3）各階級の端点の値の差を**階級の幅**という。

4）各階級に入るデータの個数を**度数**という。

5）データの個数の総数を**総度数**という。

6）各階級に度数を対応させた表を**度数分布表**という。

○度数分布をグラフとして再表現したものを**度数分布図**(histogram)又は**柱状グラフ**といいます。一般に，横軸に階級，縦軸に度数をとります。

1）柱状グラフは，階級の幅を底辺，度数を高さとする長方形を順々にかいて度数の分布を表したものです。

2）柱状グラフにおいて，両端に度数 0 の階級があるものと考えて，各長方形の上辺の中点を順に結んでできる折れ線グラフを**度数折れ線**といいます。

（※度数折れ線では，折れ線と横軸で囲まれた図形の面積は，柱状グラフの長方形の面積の総和に等しくなります。）

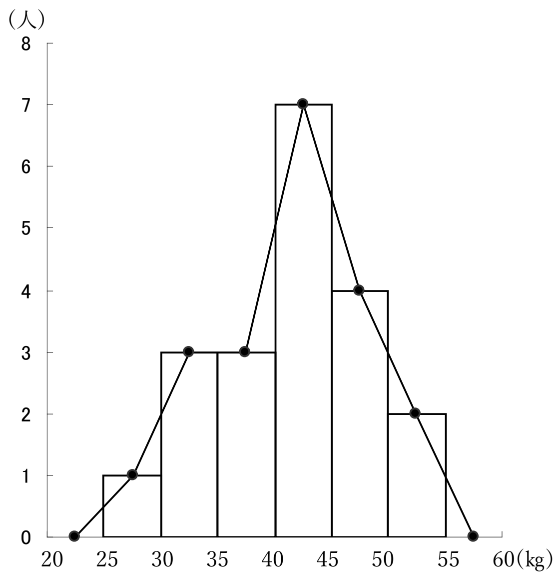
例題　次のデータは，ある学校の 1 年 B 組 20 人の生徒の握力を測定した結果である。25kg 以上 30kg 未満を階級の一つとして，**階級の幅**が 5kg の度数分布表を作り，柱状グラフと度数折れ線をかけ。（単位：kg）

32.5　40.6　47.2　36.2　50.6　49.7　45.5　39.1　48.1　42.8
38.5　53.2　34.7　44.3　28.5　42.1　41.2　43.1　30.5　40.1

［解答］(1) **度数分布表**

階級(kg) 以上～未満	**階級値** (kg)	**度数** (人)
25～30	27.5	1
30～35	32.5	3
35～40	37.5	3
40～45	42.5	7
45～50	47.5	4
50～55	52.5	2
総度数		20

(2) 柱状グラフと度数折れ線



問 3.6　次のデータは，ある学校の 50 人の学生についての 1 分間あたりの脈拍数を測定したものである。50 回以上 55 回未満を階級の 1 つとして，階級の幅が 5 回の度数分布表をつくり，柱状グラフと度数折れ線をかけ。

60　75　70　70　72　52　68　65　71　62　56　58　82　64　66　55　67
73　65　77　75　54　69　67　86　70　71　76　61　64　80　77　61　62
63　71　73　80　67　68　71　68　70　72　78　69　69　84　72　79

□□□□【参考】□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

欠測についてお話します。

欠測て何ですか？

統計調査において，何らかの理由によりデータを得ることができない場合があります。これを**欠測**と言います。

どんなときに欠測が起こりますか？

(1) ある調査項目の質問に対する回答忘れや回答拒否等による欠測を**無回答**といいます。
(2) 一定の期間を置いて調査を繰り返す際に，特定の調査回以降の回答が得られなかったような欠測を**脱落**といいます。
(3) 一定の調査期間が経過した時点で調査を終了した場合，その時点以降のデータは欠測となり，この欠測を**打切り**といいます。
(4) 調査期間中に，調査対象者の一部が除外されたことに伴うデータの欠測を**切断**といいます。

（※調査期間中に，調査項目の一部が除外されたことに伴うデータの欠測は打切りです。）

ここで一言、注意を！

重要　統計情報を見るとき，必ず表面上に現れていないデータ(欠測)があることを認識しておくことが必要です。
皆さんは，示された統計情報だけを単に見る(信用する)のではなく、その統計情報がどのようなデータに基づいて作成されたものなのかまで気を付けて見る(信憑性を疑う)習慣を身に付けてください。

□□