前回のQ＆A

Q1) ダミー変数とは何ですか？

A1) 複数のカテゴリ値を定義域とする１つの変数があるとき，その情報は｛0，1｝のみを定義域とするカテゴリ値－１個の変数で置き換えることができます。このような，｛0，1｝のみを定義域とする変数群をダミー変数と呼びます。

Q2) 説明が複雑すぎてわかりにくいので事例を使って欲しい（同様の要望多数）

A2) 十分に事例を説明するには，どうしても時間が足りないのですが，そうするように努力します。

統計学第４回　データを１つの値にまとめる（代表値）

（１）２つの戦略

- データを１つの値にまとめるとは，分布の特徴を１つの値で代表させる，ということである。このような値を，代表値と呼ぶ。代表値は，記述統計量 (descriptive statistics) の１つである。
- 代表値を求める単純な戦略としては，誰でも思いつくだろうが，２つが考えられる。分布の位置と，分布の広がりである。例えば，正規分布だったら，$N(\mu, \sigma^2)$ という形で表されるように，平均 μ，分散 σ^2 という２つの値によって分布が決まるわけだが，この場合，μ が分布の位置を決める情報で，σ^2 が分布の広がりを決める情報である。
- 一般に，調査データは，仮想的な母集団からの標本（サンプル）[1] と考えられ，データから計算される代表値は，母集団での分布の形を推定するために使われる。その意味で，これらの代表値は母数（パラメータ）と呼ばれる。
- 分布の位置を示す代表値は central tendency（中心傾向）と呼ばれ，分布の広がりを示す代表値は variability（ばらつき）と呼ばれる。
- 今回の講義で以下用いる例題は，Grimm LG (1993) Statistical Applications for the Behavioral Sciences. John Wiley & Sons, New York. の第３章と第４章からの引用が多い。代表値のような基礎的なことについてきちんと説明された教科書は意外に少ない中で，Grimm の本は丁寧に書かれていて，名著といってよい。

（２）Central Tendency

- 平均値 (mean)：平均値は，分布の位置を示す指標として，もっとも頻繁に用いられる。実験的仮説検証のためにデザインされた式の中でも，頻繁に用いられる。記述的な指標の１つとして，平均値は，いくつかの利点と欠点をもっている。普通に平均をとる (average) といって計算するのと同じことだが，数式で書くと以下の通り。
- 母集団の平均値 μ（ミューと発音する）は，

$$\mu = \frac{\sum X}{N}$$

である。X はその分布における個々の値であり，N は値の総数である。$\sum$（シグマと発音する）は，一群の値の和を求める記号である。すなわち，$\sum X = X_1 + X_2 + X_3 + ... + X_N$。

- 標本についての平均値を求める式も，母集団についての式と同一である。ただし，数式で使う記号が若干異なっている。標本平均 $\bar{X}$（エックスバーと発音する）は，

$$\bar{X} = \frac{\sum X}{n}$$

である。n は，もちろん標本数である。

例題１　値が｛5, 8, 10, 11, 12｝である母集団の平均値はいくらか？[2]

- ここで取り上げる Central Tendency には，平均値の他に，あと２つ，中央値 (median) と最頻値 (mode) がある。どれも分布の中心の位置がどの辺りかを説明するものだが，中心性 (centrality) へのアプローチが異なっている。
- 平均値は，中心性を示すために，どんなやり方をとっているのだろうか？　たまたまその値が平均値と同じであったという希な値を除けば，各々の値は，平均値からある距離をもって存在する。言い換えると，各々の値は，平均値からある程度の量，ばらついている。ある値が平均値から離れて

[1] サンプリング理論については，統計学というよりは調査法や実験計画法の範疇になるので，時間があれば別に説明する。

[2] $\mu = (5 + 8 + 10 + 11 + 12)/5 = 9.2$ であることは，小学生でもわかるだろう。もっとも，値が５つしかない母集団などというものは想像しにくいかもしれないが。

いる程度は，単純に $X - \bar{X}$ である。この，平均からの距離を，偏差（あるいは誤差）といい，x という記号で書く。つまり，$x = X - \bar{X}$ である。次の例を見ればわかるように，偏差は正の値も負の値もとるが，その合計は 0 になるという特徴をもつ。どんな形をしたどんな平均値のどんなに標本数が多い分布だろうと，偏差の和は常に 0 である。式で書くと，$\sum x = \sum X - \bar{X} = 0$ ということである。言い方を変えると，偏差の和が 0 になるように，平均値によって調整が行われたと見ることもできる。平均値は，この意味で，分布の中心であるといえる。

例題 2　分布 A が {2,4,6,8,10} という 5 つの値をもち，分布 B が {2,4,6,8,30} という 5 つの値をもっているとき，分布 A の標本平均は 6 であるから，それぞれの値の偏差は $\{-4, -2, 0, 2, 4\}$ となり，その合計は 0 である。分布 B についても確かめよ。[3]

- 重み付き平均 (weighted mean)[4] ：重み付き平均は，各々の値にある重みをかけて合計したものを，重みの合計で割った値である。式で書くと，

$$\bar{X} = \frac{n_1(\bar{X}_1) + n_2(\bar{X}_2) + ... + n_n(\bar{X}_n)}{n_1 + n_2 + ... + n_n}$$

- １つの学校区に SAT[5] の平均値が異なる 3 つの高校があり，それぞれ 425, 470, 410 だとしよう。もし，これらの値から学校区としての SAT の平均値を求めたいときは，どうしたらよいだろうか？
  単純にこれらを足して 3 で割っていいのだろうか？
- 学校区としての SAT の平均値は，3 つの高校のどれの生徒であるかにかかわりなく，全員の得点を足して，その人数で割って得るべきものである。だとすると，単純に 3 つの値を足して 3 で割るのでは具合が悪いことになる。各高校の人数は異なるので，人数の多い高校の得点の方が，総平均には余計に寄与するだろうからだ。こういう場合は，各高校の人数をそれぞれの平均点に掛けて（つまり各高校の得点総和に戻して）足し合わせ，それを人数の和（つまり学校区全体の人数）で割れば良いことが直感的にわかるだろう。これが重み付き平均の発想である。

例題 3　SAT の平均点が {425, 470, 410} であった 3 つの高校それぞれの人数が {220 人，178 人，192 人} であったなら，この学校区の SAT の総平均は何点か？[6]

練習問題　3 つの年齢群ごとの平均血圧が下の表のように記録されているとき，すべての年齢群をプールした，血圧の総平均値を求めよ。[7]

	年齢		
	20-39	40-59	60+
収縮期血圧 (mmHg)	118	128	145
拡張期血圧 (mmHg)	70	78	82
人数	13	12	16

- 度数分布の平均：離散変数の平均であれば，度数分布を出して，各値にその度数を掛けたものの和を度数の和で割ることで計算できる。言い換えると，度数で重み付けした平均値である。

$$\mu = \frac{\sum Xf}{\sum f}$$

  という式になる。
- 平均値の欠点：平均値は，例題 2 を見ればわかるように，少数の極端な値の影響を受けやすいという欠点をもつ。１つだけ極端な値があったからといって，あまりに値がそちらに引っ張られてしまっては，分布の位置を代表する値としては具合が良くない。[8]

---

[3] 標本平均は (2+4+6+8+30)/5=10 で，それぞれの値の偏差は $\{-8, -6, -4, -2, 20\}$ となるので，確かにその合計は 0 となる。分布 B は分布 A よりも平均値が大きいことがわかる。

[4] ここでは標本数が異なる複数の平均値の総平均 (grand mean) を計算する場合について説明するが，標本数以外の重みも当然ありうる。

[5] Scholastic Aptitude Test の略。米国の高校生が受ける進学適性試験。エセイティと発音する。

[6] $(220 \times 425 + 178 \times 470 + 192 \times 410)/(220 + 178 + 192) \approx 433.69$ となる。

[7] 但し，血圧の意味合いは年齢によって変わってくるからこそ，ふつう敢えて年齢群別に平均値を出すわけだから，年齢群をプールした血圧の平均値を出すことには，あまり意味はない。ここは単なる計算練習だと思って欲しい。また，ここで述べたような意味での重み付き平均を計算する必要があるのは，集計済みの二次資料から指標値を再計算するような場合なので，生データがあれば，あまり関係ない。

[8] その１つが，実は測定ミスであったり，異質な対象だったりして，外れ値である場合もあり，その場合は平均値の計算に入れないこともある。あまり機械的にやるのは良くないが，ネイマンの外れ値の検定などという手法もある。

例題４ Mackey 下院議員が，選挙区に好景気をもたらすという公約を掲げて当選したとしよう。４年後の次の選挙のときに，彼は自分が公約を果たしたと宣伝したいとする。彼の定義によると，好景気とは，選挙区に住んでいる人たちの世帯平均収入が高くなるということである。Mackey 下院議員が最初に選出されたとき，選挙区の世帯平均収入は 20,000 ドルで，２年後，年収が 14 万ドルの世帯が１つ，彼の選挙区に転入してきたとしよう。その世帯以外の年収はまったく変わらない（14,000 ドル，18,000 ドル，20,000 ドル，22,000 ドル，26,000 ドルの５世帯[9] ）として，平均値を分布の位置の指標として使った場合に平均的な世帯収入に何が起こるか見てみよう。４年後の平均世帯収入は，$(14000 + 18000 + 20000 + 22000 + 26000 + 140000)/6 = 40000$ なので，４万ドルとなる。そこで，Mackey 下院議員は，任期中に平均世帯収入は倍増した，と嘘偽りなく選挙区の人々に報告することができるわけだ。
これは，極端に高い値が，平均値を高く押し上げてしまったという例である。分布の位置の指標としては，極端な外れ値に対してこんなに敏感であっては具合が良くない。こういう極端な値が含まれている歪んだ分布の場合には，平均値という指標は誤解を生んでしまうことになる。そこで登場するのが中央値である。

- 中央値 (median)：中央値は，全体の半分がその値より小さく，半分がその値より大きい，という意味で，分布の中央である。言い換えると，中央値は，頻度あるいは値の数に基づいて分布を２つに等分割する値である。中央値を求めるには式は使わない（決まった手続き＝アルゴリズムとして，並べ替え (sorting) は必要）。極端な外れ値の影響を受けにくい（言い換えると，外れ値に対して頑健である）。歪んだ分布に対する最も重要な central tendency の指標が中央値である。

例題５ 次の分布の中央値は何か？ {1, 4, 6, 8, 40, 50, 58, 60, 62}
→この場合，小さい方から数えても大きいほうから数えても５番目の値である 40 が中央値であることは自明である。次に小さい値である 50 との距離や次に大きい値である 8 との距離は中央値を考える際には無関係である。中央値を求めるには，値を小さい順に並べ替えて[10] ，ちょうど真中に位置する値を探せばよい。この意味で，中央値は値の順序だけに感受性をもつ（= rank sensitive である）といえる。[11]

例題６ 次の標本分布の平均値と中央値は何か？ {2, 4, 7, 9, 12, 15, 17}

例題７ 次の標本分布の平均値と中央値は何か？ {2, 4, 7, 9, 12, 15, 17, 46, 54}

- 右側に２つの極端な値を加えただけ（例題７）で，平均値は大きく変わるが（ほぼ倍増），中央値は１つ右側の値に移るだけである。ところで，値の数が奇数だったら，このように順番が真中というのは簡単に決められるが，値が偶数個だったらどうするのだろうか？

例題８ 次の分布の中央値は何か？ {4, 6, 9, 10, 11, 12}

- 中央値が 9 と 10 の間にくることは明らかである。そこで，普通は 9 と 10 を平均した 9.5 を中央値として使うことになっている。[12]

例題９ 次の分布の中央値は何か？ {7, 7, 7, 8, 8, 8, 9, 9, 10, 10}

- このように同順位の値（tie という）がある場合は難しい。順番で言えば，中央値は 8 と 8 の間に来るはずだから，8 と思うかもしれないが，厳密に考えると正解は 8.33 である。[13] なぜなら，分布の値を示す数値は，間隔の中点と考えるべきだからである。つまり，測定単位が 1 なので 8 という値は 7.5 から 8.5 という間隔の中点なのだが，8 という数値が 3 つあって中央値はそのうち 2 番目と 3 番目の 8 の間に来るので，$7.5 + (2.5/3) = 8.33$ と考えるわけだ。[14] ちょっと複雑な考え方なので，もう１つ例題を挙げよう。

---

[9] ５世帯だけの選挙区などありえないから現実性はまったくない話だが，数値計算の例なので，その点は許されよ。

[10] 値の数が少ない場合には，手作業で並べ替えを行えばよいが，大量のデータを手作業で並べ替えるのは大変である。コンピュータのプログラムに値を並べ替えさせるアルゴリズムには，単純ソート，バブルソート，シェルソート，クイックソートなどがある。

[11] 平均値は値の大きさによって変わるので，value sensitive であるといえる。

[12] もっとも，本来整数値しかとらないような値について，中央値や平均値として小数値を提示することに意味があるかどうかは問題である。例えば，例題８の分布が，ある地方の水泳プールで６日間観察したときの，１日当たりの飛び込みの回数を示すものだとしよう。中央値が 9.5 ということになると，9.5 回の飛び込みというのは何を表すのか？ 半分だけ飛び込む？ つまり実体はない，単なる指標値だということになる。同様に，世帯当たりの平均子ども数が 2.4 人とかいうとき，0.4 人の子どもは実体としてはありえない。しかし，分布の位置を示す指標としては有用なので，便宜的に使っているのである。

[13] Grimm の教科書では 8.17 となっているが，tie がない場合からの自然な拡張を考えると 8.33 が正しい。なお，R や Excel では tie がある場合のこの特別な扱いをしないので，median() 関数で計算させると 8 が得られる。SPSS や SAS では確かめていないので不明だが，一般にパッケージでは，中央値に関してここで説明したほど厳密な扱いはしない。

[14] 中央値が 8 より大きな値になることについて直感的に理解するには，別の考え方もできる。順位和が同順位がない場合と等しくなるように順位をつけると，これらの順位は {2, 2, 2, 5, 5, 5, 7.5, 7.5, 9.5, 9.5} となり，中央値は 5.5 位のはずなので，5 位である 8 よりもちょっとだけ大きくなることが納得できるだろう。

例題 10 次の分布の中央値は何か？　{7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 10, 10}[15]

- さて，もう 1 歩進めて，度数分布表から中央値を計算する場合を考えてみよう。ちょっと複雑だが，理解するのは難しくない。下表は，数値を年齢階級ごとの人数の分布であり，これから年齢の中央値を求める方法を考えることにする。

年齢階級	度数	累積度数
45-49	1	76
40-44	2	75
35-39	3	73
30-34	6	70
25-29	8	64
20-24	17	56
15-19	**26**	**39**
10-14	11	13
5-9	2	2
0-4	0	0

- まず，累積度数の最大の数をみる（つまり総数をみる）。この例では 76 である。中央値の順位は $(76+1)/2 = 38.5$ 位となる。[16] 38.5 番目の値を含む年齢階級を探すと，15-19 である。その下の階級までで 13 位に到達しているので，38.5 位になるには，あと 25.5 位上がればよい。15-19 の階級には 26 個の値が含まれているので，この階級に含まれる人たちの年齢がその範囲に平均して散らばっていると仮定すれば，この階級の下限から，その間隔の $25.5/26 = 0.98$ 倍だけ積み上げればよいことになる。この表での年齢の測定単位は 1 歳だから，15-19 という年齢階級の正確な年齢の下限は 14.5 歳，上限は 19.5 歳であり，間隔は 5 歳である。したがって，中央値は，$14.5 + 5 \times 0.98 = 19.4$ 歳となる。
- ここで述べたやり方は，中央値が正確な分布の中央（少なくともその近似）になっているという特性を強めるものである。式で書けば，中央値は，

$$L + \left[\frac{(N+1)/2 - F}{f_m} \cdot h \right]$$

となる。ここで，L は中央順位を含む階級の正確な下限，F は中央順位を含む階級より下の値の総度数，f_m は中央順位を含む階級の度数，h は階級幅である。[17] ちなみに，例題 9 をこのやり方で計算すると，$7.5 + (5.5 - 3)/3 \times 1 = 8.33$ となる。
- 最頻値 (Mode)：残る最頻値は，きわめて単純である。もっとも度数が多い値を探すだけである。もっとも数が多い値が，もっとも典型的だと考えるわけである。データを見ると，最頻値が 2 つある場合があり，この場合は分布が二峰性 (bimodal) だという。[18] すべての値の出現頻度が等しい場合は，最頻値は存在しない。
- 分布の形によって，平均値，中央値，最頻値の関係は変わってくる。歪んでいない分布ならば，ばらつきの程度によらず，これら 3 つの値は一致する。二峰性だと最頻値は 2 つに分かれるが，平均値と中央値はその間に入るのが普通である。左すそを引いた分布では，平均値が最も小さく，中央値が次で，最頻値が最も大きくなる。右すそを引いた分布では逆になる。
- 平均値は，(1) 分布のすべての値を考慮した値である，(2) 同じ母集団からサンプリングを繰り返した場合に一定の値となる，(3) 多くの統計量や検定で使われている，という特長をもつ。標本調査値から母集団の因果関係を推論したい場合に，もっとも普通に使われる。しかし，(1) 極端な外れ値の影響を受けやすい，(2) 打ち切りのある分布では代表性を失う場合がある[19]，という欠点があり，外れ値があったり打ち切りがあったりする分布では位置の指標として中央値の方が優れている。

[15] 正解は 8.2 である（但し Grimm 流では 8.1）。自分で確かめてみて欲しい。

[16] Grimm の教科書だと 76 を 2 で割って 38 番目の値が中央値であると書かれているが，明らかに間違いである。もし総数を 2 で割った順位の値が中央値だとすると，例題 8 の答えが下から 3 番目で 9 ということになってしまう。総数に 1 を加えて 2 で割らなくてはいけない。

[17] この式も Grimm は間違っている。

[18] しかし隣り合う 2 つの値がともに最頻値である場合は二峰性だとはいわず，離れた 2 つの値が最頻値あるいはそれに近い場合，つまり度数分布やヒストグラムの山が 2 つある場合に，分布が二峰性だといい，2 つの異なる分布が混ざっていると考えるのが普通である。

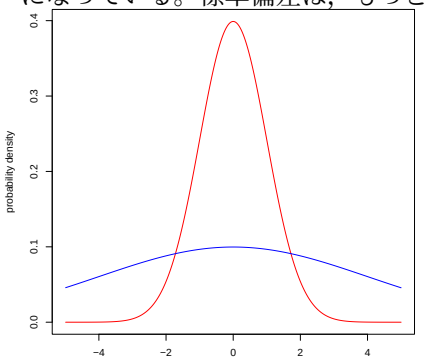
[19] 氷水で痛みがとれるまでにかかる時間とか，年収とか。無限に観察を続けるわけにはいかないし，年収は下限がゼロで上限はビル・ゲイツのそれのように極端に高い値があるから右すそを長く引いた分布になる。平均年収を出している統計表を見るときは注意が必要である。年収の平均的な水準は中央値で表示されるべきである。

最頻値は，標本をとったときの偶然性の影響を受けやすいし，もっとも頻度が高い値以外の情報はまったく使われない。しかし，試験の点で何点の人が多かったかを見たい場合は最頻値が役に立つし，名義尺度については最頻値しか使えない。

- ここで上げた 3 つの他に，幾何平均 (geometric mean) や調和平均 (harmonic mean) も，分布の位置の指標として使われることがある。幾何平均はデータの積の累乗根（対数をとって平均値を出して元に戻したもの），調和平均はデータの逆数の平均値の逆数であり，どちらもゼロを含むデータには使えない。大きな外れ値の影響を受けにくいという利点があり，幾何平均は，とくにデータの分布が対数正規分布に近い場合によく用いられる。

（3）Variability

- 分布を特徴付けるには，分布の位置だけではなく，分布の広がり具合の情報も必要である。例えば下の図の 2 つの分布は，どちらも平均 0 の正規分布なので中央値も最頻値も共通だが，幅が狭い方が標準偏差 1，幅が広い方が標準偏差 4 と，標準偏差が大きく異なるために，まったく違った外見になっている。標準偏差は，もっとも良く使われる分布の広がり具合の指標である。



- 広がり具合を示す指標は，ばらつき (variability) と総称される。ばらつきの指標には，範囲，四分位範囲，四分位偏差，平均偏差，分散（及び不偏分散），標準偏差（及び不偏標準偏差）がある。
- 範囲 (range)：最も単純なばらつきの尺度である。値のとる全範囲そのものである。つまり，最大値から最小値を引いた値になる。

例題 11　次の分布の範囲はいくらか？　{17, 23, 42, 44, 50} [20]

- ばらつきの尺度として範囲を使うには，若干の問題が生じる場合がある。極端な外れ値の影響をダイレクトに受けてしまうのである。次の例を考えてみよう。

例題 12　次の分布の範囲はいくらか？　{2, 4, 5, 7, 34}

- $34 - 2 = 32$ なのだが，2, 4, 5, 7 というきわめて近い値 4 つと，かけ離れて大きい 34 という値からなるのに，32 という範囲は，全体のばらつきが大きいかのような誤った印象を与えてしまう。ばらつきの指標としては，分布の端の極端な値の影響を受けにくい値の方がよい。
- 四分位範囲 (Inter-Quartile Range; IQR)：そこで登場するのが四分位範囲である。その前に，分位数について説明しよう。値を小さい方から順番に並べ替えて，4 つの等しい数の群に分けたときの 1/4, 2/4, 3/4 にあたる値を，四分位数 (quartile) という。1/4 の点が第 1 四分位，3/4 の点が第 3 四分位である（つまり全体の 25 %の値が第 1 四分位より小さく，全体の 75 %の値が第 3 四分位より小さい）。2/4 の点というのは，ちょうど順番が真中ということだから，第 2 四分位は中央値に等しい。[21] これを一般化して，値を小さい方から順番に並べ替えて，同数の群に区切る点を分位数 (quantile) という。百等分した場合を，とくにパーセンタイル (percentile) という。言い換えると，第 1 四分位は 25 パーセンタイル，第 3 四分位は 75 パーセンタイルである。四分位範囲とは，第 3 四分位と第 1 四分位の間隔である。パーセンタイルでいえば，75 パーセンタイルと 25 パーセンタイルの間隔である。上と下の極端な値を排除して，全体の中央付近の 50 %（つまり代表性が高いと考えられる半数）が含まれる範囲を示すことができる。
- 四分位偏差 (Semi Inter-Quartile Range; SIQR)：四分位範囲を 2 で割った値を四分位偏差と呼ぶ。もし分布が左右対称型の正規分布であれば，中央値マイナス四分位偏差から中央値プラス四分位偏差までの幅に全データの半分が含まれるという意味で，四分位偏差は重要な指標である。IQR も SIQR も少数の極端な外れ値の影響を受けにくいし，分布が歪んでいても使える指標である。

[20] いうまでもなく，$50 - 17 = 33$ である。

[21] ちょっと考えればわかるように，ちょうど 4 等分などできない場合がもちろんあって，上から数えた場合と下から数えた場合で四分位数がずれる可能性があるが，その場合はそれらを平均するのが普通である。また，最小値，最大値に，第 1 四分位，第 3 四分位と中央値を加えた 5 つの値を五数要約値と呼ぶことがある。第 1 四分位，第 2 四分位，第 3 四分位は，それぞれ Q1, Q2, Q3 と略記することがある。

例題 13 パプアニューギニアのある村で成人男性 28 人の体重を量ったところ，{50.5, 58.0, 47.5, 53.0, 54.5, 61.0, 56.5, 65.5, 56.0, 53.0, 54.0, 56.0, 51.0, 59.0, 44.0, 53.0, 62.5, 55.0, 64.5, 55.0, 67.0, 70.5, 46.5, 63.0, 51.0, 44.5, 57.5, 64.0}（単位は kg）という結果が得られた。このデータから，四分位範囲と四分位偏差を求めよ。[22]

- 平均偏差 (mean deviation)：偏差の絶対値の平均値を平均偏差と呼ぶ。四分位範囲や四分位偏差は，全データのうちの限られた情報しか使わないので，分布のばらつきを正しく反映しない可能性がある。そこで，すべてのデータを使ってばらつきを表す方法を考えよう。すべての生の値は，平均値からある距離をもって分布している。この距離は既に述べたように偏差あるいは誤差と呼ばれる。偏差の大きさは，分布のばらつきを反映している。

例題 14 分布 A が {11, 12, 13, 14, 15, 16, 17}，分布 B が {5, 8, 11, 14, 17, 20, 23} だとする。どちらも平均値は 14 である。しかし，分布 B は分布 A よりもばらつきが大きい。言い換えると，分布 B の方が分布 A よりも平均値からの距離が大きい。しかし，それをどうやって 1 つの値として表すことができるだろうか？　ただ合計しただけでは，平均値のところで述べたように，偏差の総和は必ずゼロになってしまう。これはマイナス側の偏差がプラス側の偏差と打ち消しあってしまうためなので，偏差の絶対値の総和を出してやればいいというのが最も単純な発想である。それだけだとサンプル数が多いほど大きくなってしまうので，値 1 つあたりの偏差の絶対値を出してやるためにサンプル数で割ることが考えられる。これが平均偏差の考え方である。すなわち，平均偏差 MD は，

$$MD = \frac{\sum |X - \mu|}{N}$$

で定義される。μ は平均値，N はサンプル数である。

- 平均偏差はすべてのデータを使い，かつ少数の外れ値の影響は受けにくいという利点があるが，絶対値を使うために他の統計量との数学的な関係がなく，標本データから母集団統計量を推定するのに使えないという欠点がある。
- 分散 (variance)：マイナス側の偏差とプラス側の偏差を同等に扱うためには，絶対値にするかわりに二乗しても良い。つまり，偏差の二乗和の平均をとるわけである。これが分散という値になる。分散 V は，

$$V = \frac{\sum (X - \mu)^2}{N}$$

で定義される。[23] 標本数 n で割る代わりに自由度 $n-1$ で割って，不偏分散 (unbiased variance) という値にすると，標本データから母集団の分散を推定するのに使える。即ち，不偏分散 V_{ub} は，

$$V_{ub} = \frac{\sum (X - \bar{X})^2}{(n-1)}$$

である。

- 標準偏差 (standard deviation)：分散では大きすぎる値になるので，その平方根をとったものが標準偏差である。不偏分散の平方根をとったものは，不偏標準偏差となる。もし分布が正規分布ならば，Mean±2SD[24] の範囲にデータの 95 %が含まれるという意味で，標準偏差は便利な指標である。

練習問題 例題 14 の 2 つの分布について，平均偏差，分散，不偏分散，標準偏差，不偏標準偏差を計算せよ。

- 標準誤差 (standard error) と変動係数 (coefficient of variation)：生データの分布のばらつきの指標ではないが，関連するのでここで示しておく。不偏標準偏差を $\sqrt{N}$ で割った値は，平均値の推定幅を示す値となり[25]，標準誤差 (standard error) として知られている。SD と SE は論文などでは良く混用されているが，意味がまったく違う。また，標準偏差（不偏標準偏差ではない）を平均値で割って 100 を掛けた値を変動係数という。即ち，平均値に対して，全測定値が何%ばらついているかを示す，相対的なばらつきの指標である。これは測定誤差を示すときなどに使われる値であり，母集団統計量である。

（4）まとめ

- データの分布は，位置とばらつきを示す 2 つの代表値に集約して示すのが普通である。分布に外れ値が多い，歪みが大きい，尺度水準が低い，など分布を仮定できない場合は，中央値と四分位偏差を用い，そうでない場合は平均値と（不偏）標準偏差を用いて，位置±ばらつき，という形で示すのが普通である。

---

[22] R の fivenum 関数を使うと，Q1=52.00，Q2=55.50，Q3=61.75 とわかる。これより，四分位範囲は $Q3 - Q1 = 9.75$，四分位偏差はそれを 2 で割って 4.975 である。

[23] 実際に計算するときは 2 乗の平均から平均の 2 乗を引くと簡単である。

[24] 普通このように 2SD と書かれるが，正規分布の 97.5 パーセント点は 1.959964... なので，この 2 は，だいたい 2 くらいという意味である。

[25] 平均値の分散は生データの分散の 1/N になることと，N が大きいとき，元の分布によらず平均値は正規分布に近づく（中心極限定理）ため。