2014 年度春学期　統計学　第５回
分布をまとめる ー 記述統計量（平均・分散など）

代表値

度数分布のヒストグラムによる表現は，視覚的にはよくわかる表現です。しかし，度数分布から取り出される情報を今後の処理に用いたり，比較したりするには，分布を 1 つの数字で表現する必要があります。これを，**代表値**といいます。ここでは，もっともよく使われる代表値である算術平均と，分布を表現するもうひとつの重要な指標である分散，さらに分散を発展させた「モーメント」の考えについて説明します。

算術平均

データを $x_1, x_2, \ldots, x_n$，総データ数を n とするとき，**算術平均 (arithmetic mean，相加平均ともいう)** は次の式で定義されます。

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

つまり，**算術平均 ＝（データの合計）／（データ数）**です。ふつう，単に「平均」といえば算術平均のことをさします。

度数分布から算術平均を求める

データの度数分布がわかっているときに，その平均を求めるにはどうすればよいでしょうか？ 平均とはデータの合計をデータの個数で割ったものです。一方，ある階級の度数は「その階級値をとるデータが，何個あるか」を表しています。そこで，

平均＝（データの合計）／（データ数）
＝（［階級値×度数］の合計）／（データ数）
＝［階級値×（度数／データ数）］の合計
＝［階級値×相対度数］の合計

ですから，**「平均 ＝［階級値×相対度数］の合計」**ということになります。

分散と標準偏差

分布をもっとも簡単に 1 つの数字で表したのが代表値ですが，代表値だけでは，その分布が「どのくらいばらついているか」は表現できません。その例を見てみましょう。つぎのようなデータの組 A, B, C があるとします。

A: 0, 3, 3, 5, 5, 5, 5, 7, 7, 10
B: 0, 1, 2, 3, 5, 5, 7, 8, 9, 10
C: 3, 4, 4, 5, 5, 5, 5, 6, 6, 7

これらの平均はいずれも5で，平均値ではこれらの分布を区別して表現することはできません。これらの分布の違いは，ばらつきにあります。

AとBは分布の幅（レンジ）は違いませんが，分布の平均値への集まり具合がちがいます。レンジは分布の両端の値しか使っていないので分布の平均値への集まり具合を表現することはできませんが，次に述べる**分散**や**標準偏差**は，分布内のすべてのデータを使うので，集まり具合を表現できます。

各データと平均との差を**偏差**といい，各データが平均からどのくらい離れているかを表します。「偏差の平均」を求めれば，このデータ組の「データの平均からの散らばり具合」がわかりそうですが，平均値はデータ組のちょうど真ん中の値ですから，「偏差の平均」は0になってしまいます。

そこで，「偏差の平均」のかわりに「(偏差)^{2}の平均」を用います。(偏差)^{2}はすべて正ですから，「(偏差)^{2}の平均」，すなわち「各データについての偏差の２乗の合計を総データ数で割ったもの」でばらつきの程度を表現できます。これが**分散 (variance)** です。式で書くと，各データを $x_1, x_2, \ldots, x_n$，総データ数を n，平均を $\bar{x}$ とするとき，分散 σ^2 はつぎのようになります。

$$
\begin{aligned}
\sigma^2 &= \frac{1}{n}\left\{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2
\end{aligned}
\tag{2}
$$

また，分散の平方根を**標準偏差 (standard deviation, SD)** といいます。データの単位がm（メートル）のとき，分散の単位はm^2，すなわち平方メートルになってしまいますが，標準偏差の単位は同じmです。

分散を求めるとき，なぜ偏差の絶対値をとらずに偏差を２乗するのか？

確かに，偏差の絶対値を使って計算しても，「偏差を全部正の値にしてから平均する」という目的は達せられます。しかし，絶対値の計算は２乗よりも簡単そうですが，実はそうではありません。２乗の計算は，どんな数に対しても同じ手続きでできますが，絶対値の計算は，正の数と負の数とで別の手続きが必要です。みなさんも，高校の数学の時間に，「$y = 2x + 3$ のグラフを描け」といった問題で，ややこしい場合分けをやった記憶があると思います。こういう事情で，偏差の絶対値の平均は用いられず，偏差の２乗の平均である分散が用いられているのです。さらに，２乗を考えると，これを３乗，４乗，…に発展させることができます。これについては，すぐあとで説明します。

度数分布から分散を求める

上で，度数分布から平均を求める方法として「平均 ＝［階級値×相対度数］の合計」となることを示しました。分散は**「(偏差）^{2} の平均」**ですから，上の計算を利用すると，**「分散＝［(偏差）^{2}× 相対度数］の合計」**すなわち**「分散＝［(階級値 − 平均）^{2}× 相対度数］の合計」**という計算で求められます。

モーメント

度数分布において，分布しているデータを変数 X で代表し[1]，ある階級の階級値を x で表します。また，階級値が x である階級の相対度数を，$f(x)$ で表すことにします。このとき，平均を $E(X)$ で表すこ

[1]つまり，分布そのものをひとつの変数 X であらわしていることになります。この考え方は，次回の講義で「確率変数」を説明するときに，もう一度出てきます。

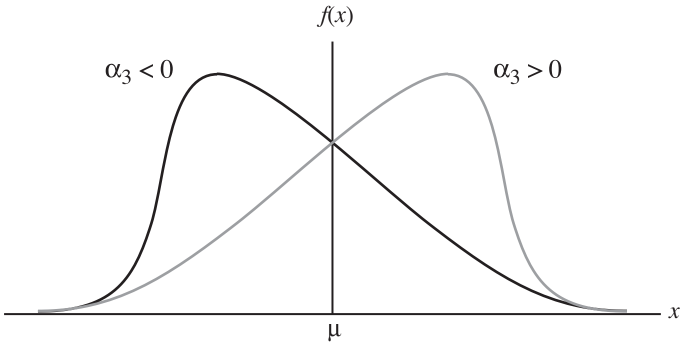

図 1: 歪度（ヒストグラムの上辺を連続曲線で表示）

とにすると，上で示した，度数分布から平均を求める計算により

$$E(X) = \sum_x x f(x) \tag{3}$$

となります。$E(X)$ すなわち平均を μ で表すことにします。

分散は，(偏差)^2，すなわち $(X-\mu)^2$ の平均ですから，(3) 式と同様の書き方を用いれば

$$V(X) = E((X-\mu)^2) = \sum_x (x-\mu)^2 f(x) \tag{4}$$

と表すことができます。$V(X)$，すなわち分散は，σ^2 で表すこともよくあります。こうすると，標準偏差は σ ということになります。

これらを一般的に表して，「変数 X の関数 $g(X)$ の平均」$E(g(X))$ を考えると，

$$E(g(X)) = \sum_x g(x) f(x) \tag{5}$$

となり，上の平均や分散は (5) 式の特殊な場合と考えることができます。

$E(g(X))$ の別の特殊な場合として，$E(X^k)$ や $E((X-\mu)^k)$ (k は自然数) を考えます。これらを X の k 次の**モーメント (積率)** とよびます。$E(X^k)$ を原点のまわりのモーメントとよんで μ'_k で表し，$E((X-\mu)^k)$ を平均のまわりのモーメントとよんで μ_k で表します。平均 μ は，実は原点のまわりの 1 次モーメント μ'_1 であり，分散 $V(X)$ は平均のまわりの 2 次のモーメント μ_2 であるということになります。

「モーメント」という名前は，力学の用語からの類推から来ています。力学では，「物体中の各点の原点（あるいは重心）からの距離 × その点にある質量（あるいは働く力）」を物体中の全ての点について合計したものを，「原点（重心）のまわりの 1 次のモーメント」といいます。$E(X)$ を求める式で，x を距離，$f(x)$ を質量（力）とすれば力学でのモーメントと同じになります。

平均や分散は，分布の特徴を記述するのにもっとも頻繁に使われる量です。さらに高次のモーメントを用いると，分布の特徴をより細かく記述できます。その中でよく使われるのは，$\alpha_3 = \mu_3/\sigma_3$ で定義される**歪度 (skewness)** と，$\alpha_4 = \mu_4/\sigma_4$ を用いて $\alpha_4 - 3$ で定義される**尖度 (kurtosis)** です[2]。

$(X-\mu)^3$ は，$x > \mu$，すなわちデータが平均より大きいときは正で，$x < \mu$ のときは負になります。したがって，データが平均より大きい階級において相対度数 $f(x)$ が大きければ，μ_3 は正になり，データ

[2] "3" は，正規分布モデル（第 2 部で説明します）の場合の α^4 の値です

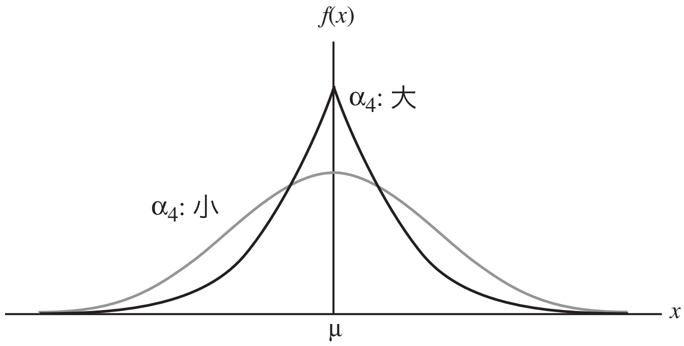

図 2: 尖度（同上）

が平均より小さい階級で相対度数 $f(x)$ が大きければ μ_3 は負になりますから，歪度は，$f(x)$ のヒストグラムの，正負の方向への偏り具合をあらわします。

また，$(X-\mu)^4$ は，データが平均に近いとき非常に小さくなりますから，単峰性分布（ヒストグラムの峰がひとつである分布）の場合に μ_4 の値が大きくなるためには，$f(x)$ が $x=\mu$ 付近で突出して大きくなる必要があります。すなわち，尖度が大きいことは，$f(x)$ のヒストグラムが，μ 付近で上にとがっていることを示しています。

標準得点

ある人が，数学の試験で 100 点満点で 70 点をとったとします。70 点という点数そのものには，問題全体の 70%に正答したので，その試験についてはまあまあの出来，という意味はもちろんあります。しかし，大学などの受験においては，同じ試験を受けたすべての受験生の中で上位に入らなければ合格はできませんから，その 70 点という点数が「同じ試験を受けた他の受験生に比べて，上位なのか下位なのか」を知ることが重要です。同じ 70 点でも，他の受験生が皆 50 点そこそこなら，70 点をとった人は上位に位置するでしょうし，他の受験生が皆 90 点以上なら，他の人より大幅に劣っていることになって，意味合いは全然違います。

このように，あるデータが分布の中でどのぐらいの位置にいるかを数値で表現するために，「そのデータが，分布の平均に比べて，標準偏差の何倍上回っているか（あるいは下回っているか）」を求めます。この値を**標準得点**といいます。たとえば，「あるデータを標準得点に換算すると 1.0 である」ということは，そのデータが平均にくらべて標準偏差の 1.0 倍上回っていることを意味しています。また，標準得点が -1.5 なら，平均にくらべて標準偏差の 1.5 倍下回ったであることを意味しています。

標準得点を求めるために，分布を平均 0・標準偏差が 1 になるように「変換」することを考えます。「分布を変換する」とは，分布に含まれる各々のデータについて，一斉に同じ計算を施して，別の分布を作ることです。このような計算をしたとき，あるデータが変換の結果 1.0 になったとすれば，それは平均 0・標準偏差 1 である分布においての 1.0 という値ですから，これは平均よりも 1.0 倍上回っていることを示しており，標準得点に変換されていることになります。

どういう計算をすれば，分布を平均 0・標準偏差が 1 になるように「変換」することができるでしょうか？ 例えば，分布に含まれる各データについて一斉に「10 を引く」という計算をすると，どのデー

タも 10 小さくなるわけですから，平均も 10 小さくなることは容易に理解できます。そこで，10 を引くかわりに「元の平均を引く」という計算をすると，平均は「(元の平均) ― (元の平均)」，すなわち 0 になります。

標準偏差のほうはどうでしょうか。いま述べた「平均を 0 にする変換」をしたあと，各データを一斉に「2 倍する」という計算をするとしましょう。各データが 2 倍になったとき，平均は「データの合計／総データ数」ですから，平均も 2 倍になりますが，いまは平均は 0 なので，2 倍してもやはり 0 です。そうすると，各データと平均との差である偏差が 2 倍になります。したがって，分散は「偏差の 2 乗の平均」ですから，分散は 2 の 2 乗，すなわち 4 倍になります。標準偏差は分散の平方根ですから，4 倍になったものの平方根で，やはり 2 倍になります。そこで，2 倍するかわりに「(1/元の標準偏差) 倍する」という計算をすると，標準偏差は「(元の標準偏差) の (1/元の標準偏差) 倍」で，すなわち 1 になります。

以上のことから，どんな分布でも「平均を引いて，標準偏差で割る」という計算をすれば，平均 0・標準偏差 1 である分布に変換されます。したがって，あるデータに「そのデータが含まれる分布の平均を引いて，標準偏差で割る」という計算をすると，標準得点に変換できます。

この変換を，図 3 でヒストグラムを使って説明しています。ここでは，元の平均を μ，元の標準偏差を σ で表しています。ヒストグラムにおいて，度数を表しているのは柱の高さではなく面積であるため，「$(1/\sigma)$ 倍する」という変換でヒストグラムの横方向の広がりが変わると，それに応じて高さも変わることに注意してください。

より一般的に，分布中のデータにすべて同じ定数 a をかけて，さらにすべて同じ定数 b を加えることを考えます。この計算を式で書くと，定数 a, b をもってきて，もとの分布の各データ x_i に対して $z_i = ax_i + b$ という計算をして，別のデータ z_i を作ることになります。これを，「各データ x_i を，$z_i = ax_i + b$ という 1 次式で z_i に変換する」といいます。また，分布そのものを代表して X であらわして，「分布 X を，$Z = aX + b$ という 1 次式で分布 Z に変換する」という言い方をすることもあります。

このとき，変換前の分布の平均を μ_x・分散を σ_x^2・標準偏差を σ_x とし，変換後の分布の平均を μ_z・分散を σ_z^2・標準偏差を σ_z とすると，

$$\mu_z = a\mu_x + b, \quad \sigma_z^2 = a^2\sigma_x^2, \quad \sigma_z = |a|\sigma_x \tag{6}$$

となります。計算は，付録に載せています。標準得点に変換する計算は，$a = \frac{1}{\sigma_x}, b = -\frac{\mu_x}{\sigma_x}$ とおいた場合に相当し，このとき

$$\mu_z = \frac{1}{\sigma_x}\mu_x + \left(-\frac{\mu_x}{\sigma_x}\right) = 0, \quad \sigma_z = |\frac{1}{\sigma_x}|\sigma_x = 1 \tag{7}$$

となるので，この計算で新しい分布をつくると，その平均は 0，標準偏差は 1 となります。

標準得点（平均 0，標準偏差 1）に対して，さらに $a = 10$, $b = 50$ とおいて各データをもう一度変換してみます。すると，分布を変換する (6) 式に $\mu_x = 0, \sigma_x = 1, a = 10$, $b = 50$ を代入すると分かるように，変換後の分布は平均 50 点，標準偏差 10 点となります。このように各データを変換して得られる得点が，受験でおなじみの偏差値です。例えば，偏差値 70 点とは，その試験の平均点よりも標準偏差の 2 倍だけ高い点数であることを表しています。これは，学力テストは 100 点満点で行われることが多いため，30 点〜70 点あたりのなじみのある値を使って分布中の位置を表現するために考案されたものです。

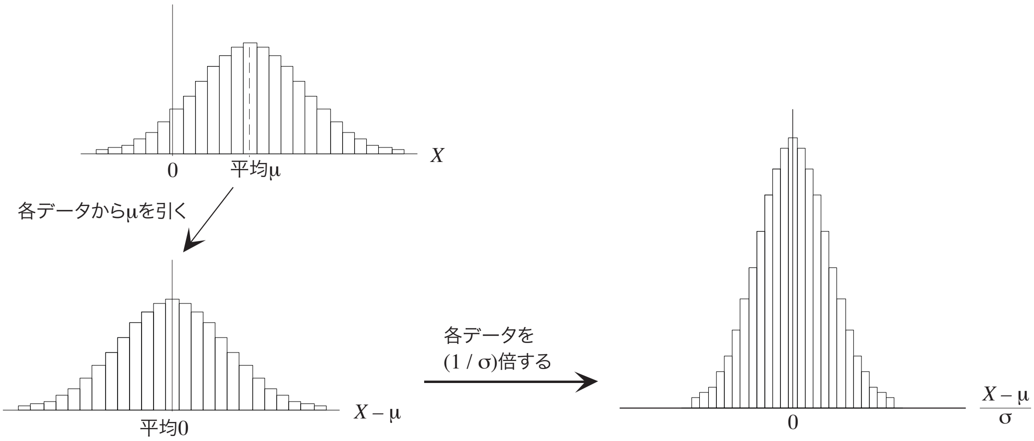

図 3: 度数分布の変換

今日の演習

１．下の度数分布表について，表の空欄を埋めて，平均・分散を求めてください．

階級	階級値	相対度数	階級値×相対度数	偏差	(偏差)^2	(偏差)^2 × 相対度数
0〜9（点）	5	0.04				
10〜19	15	0.16				
20〜29	25	0.08				
30〜39	35	0.12				
40〜49	45	0.10				
50〜59	55	0.10				
60〜69	65	0.12				
70〜79	75	0.08				
80〜89	85	0.18				
90〜100	95	0.02				
合計		1.0	＝平均			＝分散

表 1: 度数分布から平均・分散を求める

２．ある 10 人のクラスで数学の試験を行ったところ，その得点は 20, 45, 50, 50, 60, 60, 65, 70, 70, 100（点）でした．このとき，

1. 平均と標準偏差を求めてください．
2. 得点が 70 点の人の偏差値は何点ですか．
3. 偏差値 65 点は，試験の得点では何点に相当しますか．

付録：(6) 式の導出

算術平均および分散の定義から，

$$
\begin{aligned}
\mu_z &= \frac{z_1 + z_2 + \cdots + z_n}{n} \\
&= \frac{(ax_1 + b) + (ax_2 + b) + \cdots + (ax_n + b)}{n} \\
&= \frac{a(x_1 + x_2 + \cdots + x_n) + nb}{n} = a\mu_x + b
\end{aligned}
\tag{A1}
$$

となります。また，

$$
\begin{aligned}
\sigma_z^2 &= \frac{1}{n}\left\{(z_1 - \mu_z)^2 + (z_2 - \mu_z)^2 + \cdots + (z_n - \mu_z)^2\right\} \\
&= \frac{1}{n}\left\{((ax_1 + b) - (a\mu_x + b))^2 + ((ax_2 + b) - (a\mu_x + b))^2 + \cdots + ((ax_n + b) - (a\mu_x + b))^2\right\} \\
&= \frac{1}{n}\left\{a^2(x_1 - \mu_z)^2 + a^2(x_2 - \mu_z)^2 + \cdots + a^2(x_n - \mu_z)^2\right\} \\
&= a^2\frac{1}{n}\left\{(x_1 - \mu_z)^2 + (x_2 - \mu_z)^2 + \cdots + (x_n - \mu_z)^2\right\} = a^2\sigma_x^2
\end{aligned}
\tag{A2}
$$

となりますから，$\sigma_z = |a|\sigma_x$ となります。