2　平均値と分散

2.1　平均値／期待値

2.1.1　有限データの場合

定義 2.1.1 有限データ X に含まれる数値の総和をデータのサイズで割った値を、そのデータの平均値（mean value）と言い記号 $E[X]$ で表します。

数の羅列　$10, 10, 8, 7, 10, 8, 8, 6, 10, 9, 8, 9$

例えばこの例では、

$$(\text{平均値}) = \frac{10+10+8+7+10+8+8+6+10+9+8+9}{12} = \frac{103}{12} = 8.58333\ldots$$

ですが、『足す順番』はどうでも良いことに注意して（そもそも、データに含まれる数値全体に『自然な順序』が定まっているとは限りません）、同じものはまとめて足してしまいましょう。10 はを４個、9 を２個足しているわけですから、まとめると

$$(\text{平均値}) = \frac{10\cdot 4 + 9\cdot 2 + 8\cdot 4 + 7\cdot 1 + 6\cdot 1}{12}$$

となります。また、更に変形して分母のデータサイズ（12）を各度数の下にもって行って

$$(\text{平均値}) = 10\times\frac{4}{12} + 9\times\frac{2}{12} + 8\times\frac{4}{12} + 7\times\frac{1}{12} + 6\times\frac{1}{12}$$

と計算しても良く、ここに現れる分数 $\dfrac{(\text{度数})}{(\text{データサイズ})}$ は相対度数ですから

$$(\text{平均値}) = \sum (\text{数値})\times(\text{相対度数}) = \sum (\text{数値})\times(\text{確率})$$

と云う風になっている事が分かります。

2.1.2　可算無限データの場合

定義 2.1.2 可算無限種類の数値からなるデータ X に対して、各数値にその数値の相対度数を掛けた値の総和が絶対収束する場合、その総和を平均値と言い記号 $E[X]$ で表します。

サイコロを振って初めて５が出るまでの振る回数を X と置けば、X の確率分布表は

n	1	2	3	$\ldots$	n	$\ldots$
$P[X=n]$	$\frac{1}{6}$	$\frac{5}{6}\frac{1}{6}$	$\left(\frac{5}{6}\right)^2\frac{1}{6}$	$\cdots$	$\left(\frac{5}{6}\right)^{n-1}\frac{1}{6}$	$\cdots$

となりますから、平均値は

$$\frac{1}{6}\left\{1 + 2\left(\frac{5}{6}\right) + 3\left(\frac{5}{6}\right)^2 + \cdots\right\} = 6 = E[X]$$

です。１回振るごとに５が『$\frac{1}{6}$ 回出ている』と考えれば、確かに６回振ってようやく『１回出た』事になるわけですね。

この様に、データを数値の集まりとしてよりもくじ引きの結果として考える空気が濃厚な場合は確率計算によってその確率変数が『どの程度の値をとると期待されるか』を求めていると考えられます。そこで確率変数に対しては平均値と言う代わりに『期待値（expected value）』と呼ぶ事もあります。

別の例で見てみましょう。前回見た漸化式によって定まるデータ Y では：

数	2	4	8	16	32	$\cdots$	2^n	$\cdots$
相対度数	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\cdots$	$\frac{1}{2^n}$	$\cdots$

$$(\text{数値と相対度数の積の総和}) = 2\cdot\frac{1}{2} + 4\cdot\frac{1}{4} + \cdots = +\infty$$

となって $+\infty$ に発散していますから平均値は存在しません（平均値は $+\infty$ であるとする流儀もあります）。

また、負の数もとり得るようなデータでは、数値と相対度数の積の総和は負の項を含む無限級数になります。これを足して行く『順番』は考慮されないはずですが、一般に負の項を含む級数が収束はしても絶対収束しない場合、『足す順番を換えると総和が変わってしまう』場合があり、平均値を定義することは難しくなります（どの順番で足したものを平均値とするのが自然なのか？）。従って、普通、総和が絶対収束する場合のみ平均値が存在すると考えます。

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots = \log 2$$

$$1 - \frac{1}{2} - \frac{1}{4} + \frac{1}{3} - \frac{1}{6} - \frac{1}{8} + \frac{1}{5} - \frac{1}{10} - \frac{1}{12} + \frac{1}{7} - \frac{1}{14} - \frac{1}{16} + \cdots$$

$$= \left(1 - \frac{1}{2}\right) - \frac{1}{4} + \left(\frac{1}{3} - \frac{1}{6}\right) - \frac{1}{8} + \left(\frac{1}{5} - \frac{1}{10}\right) - \frac{1}{12} + \left(\frac{1}{7} - \frac{1}{14}\right) - \frac{1}{16} + \cdots = \frac{1}{2}\log 2$$

2.1.3　非可算無限データの場合

平均値とは数値とその相対度数（確率）の積の総和のことでした：

$$\text{（平均値）} = \sum_{\text{全部足す}} \text{（数値）} \times \text{（相対度数）}.$$

しかし非可算無限データでは素朴な意味での相対度数は意味を成していませんし、総和も可算個の和ではないため $\sum$ は使えません。

データ X の密度関数が $f(x)$ であるとは、任意の $a \leq b$ に対して

$$P[a \leq X \leq b] = \int_a^b f(x)dx$$

となるような関数のことでした。

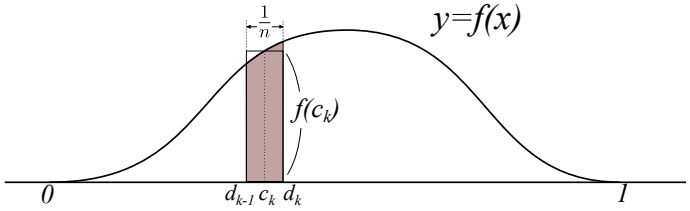
例えばデータの最小値が 0、最大値が 1 であるとき、区間 $[0,1]$ を n 等分割し、

$$a = d_0 < d_1 < \cdots < d_n = b, \quad d_k = \frac{k}{n} \quad (k = 0, 1, \ldots, n)$$

各小区間 $[d_{k-1}, d_k] = \left[\frac{k-1}{n}, \frac{k}{n}\right]$ から真ん中の点 $c_k = \frac{k-\frac{1}{2}}{n}$ を取って作られたリーマン和：

$$\sum_{k=0}^{n} c_k f(c_k) \frac{1}{n}$$

は、$n \to +\infty$ で定積分 $\int_0^1 xf(x)dx$ に収束しますが、n が十分大きければ長方形の面積 $f(c_k) \times \frac{1}{n}$ は面積 $\int_{d_{k-1}}^{d_k} f(x)dx$ にほぼ等しく、これは全データ中の d_{k-1} 以上 d_k 以下の数値の相対度数ですから、このリーマン和は各階級 $[d_{k-1}, d_k]$ に入るデータを全て階級の真ん中の値 c_k で置き換えて得られる階級分け近似データの平均値を近似計算したものと考えられます。



これを $n \to +\infty$ としたものを考えると、階級の幅を極限まで狭くして行く訳ですから、近侍の精度が高まって行き、極限値がデータ X の平均値であると考えられます。

定義 2.1.3 密度関数が $f(x)$ であるデータ／確率変数 X に対して次の積分が有限値として存在する時、これを X の平均値（期待値）と言い記号 $E[X]$ で表します：

$$\underbrace{\int_{-\infty}^{\infty}}_{\text{全部足す}} \underbrace{x}_{\text{数値}} \underbrace{f(x)dx}_{\text{相対度数}}$$

問題 2.1.4 確率変数 X の分布密度関数が次の $h(x)$ で与えられているとき、確率 $P[-0.5 \leq X \leq 1.5]$、平均値 $E[X]$ を求めて下さい。

$$h(x) = \begin{cases} \frac{3}{4}(1-x^2) & -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$P[-0.5 \leq X \leq 1.5] = \int_{-0.5}^{1.5} h(x)dx = \int_{-0.5}^{1} \frac{3}{4}(1-x^2)dx = \left[\frac{3}{4}x - \frac{1}{4}x^3\right]_{-0.5}^{1} = \frac{27}{32}$$

$$\int_{-\infty}^{\infty} xh(x)dx = \int_{-1}^{1} x\frac{3}{4}(1-x^2)dx = \left[\frac{3}{8}x^2 - \frac{3}{16}x^4\right]_{-1}^{1} = 0 = E[X]$$

□

問題 2.1.5 確率変数 X の分布密度関数が $h(x) = \frac{1}{\pi}\frac{1}{1+x^2}$ であるとき、期待値 $E[X]$ を求めて下さい。

広義積分を計算すると

$$\int_{-\infty}^{\infty} \frac{1}{\pi}\frac{x}{1+x^2}dx = \lim_{m \to -\infty, M \to +\infty} \int_m^M \frac{1}{\pi}\frac{x}{1+x^2}dx$$

$$= \lim_{m \to -\infty, M \to +\infty} \left[\frac{1}{2\pi}\log\left|1+x^2\right|\right]_m^M$$

となり、この極限値は存在しません。従って期待値は存在しません。 □

2.2　データの変換、その密度と期待値

データ X に含まれるそれぞれの数値 x に対して $F(x)$ と云う新たな数値を考え、この $F(x)$ を全て集めて１つの新たなデータ $F(X)$ として考える事があります。

2.2.1　１次関数

$F(x) = ax + b\,(a > 0)$ のときは $F(X) = aX + b$ です。X の密度が $h(x)$ であるときに $F(X)$ の密度を計算すると、

$$P[v \le F(x) \le w] = P[v \le aX + b \le w] = P\left[\frac{v-b}{a} \le X \le \frac{w-b}{a}\right] = \int_{\frac{v-b}{a}}^{\frac{w-b}{a}} h(x)dx$$

ここで $\frac{z-b}{a} = x$ と置換すれば

$$P[v \le F(X) \le w] = \int_a^b h\left(\frac{z-b}{a}\right)\frac{1}{a}dz$$

となりますから $F(X)$ の密度は $g(x) = \frac{1}{a}h\left(\frac{x-b}{a}\right)$ であることが分かりました。

そこで期待値を計算してみると

$$E[F(X)] = \int_{-\infty}^{\infty} xg(x)dx = \int_{-\infty}^{\infty} x\frac{1}{a}h\left(\frac{x-b}{a}\right)dx$$

ですから、ここで $\frac{x-b}{a} = z$ と置換すれば、$az + b = x$ となって

$$E[F(X)] = \int_{-\infty}^{\infty}(az+b)\frac{1}{a}h(z)a\,dz = \int_{-\infty}^{\infty}(az+b)h(x)dx = a\int_{-\infty}^{\infty} zh(z)dz + b\int_{-\infty}^{\infty} dz$$

$$E[aX + b] = aE[X] + b$$

が分かります（平均値の線形性）。

2.2.2　２次関数

今度は X の密度が $h(x)$ であるときに、$(X-m)^2$ の密度を計算してみましょう。$v, w \ge 0$ のとき、

$$\begin{aligned} P[v \le (X-m)^2 \le w] &= P[\sqrt{v} \le X - m \le \sqrt{w}] + P[-\sqrt{w} \le X - m \le -\sqrt{v}] \\ &= P[m + \sqrt{v} \le X \le m + \sqrt{w}] + P[m - \sqrt{w} \le X \le m - \sqrt{v}] \\ &= \int_{m+\sqrt{v}}^{m+\sqrt{w}} h(x)dx + \int_{m-\sqrt{w}}^{m-\sqrt{v}} h(x)dx \end{aligned}$$

ここで第 1 の積分では $x = m + \sqrt{z}$、第 2 の積分では $x = m - \sqrt{z}$ とおけば

$$\begin{aligned} &= \int_v^w h(m+\sqrt{z})\frac{1}{2\sqrt{z}}dz + \int_w^v h(m-\sqrt{z})\left(-\frac{1}{2\sqrt{z}}\right)dz \\ &= \int_v^w \frac{h(m+\sqrt{z}) + h(m-\sqrt{z})}{2\sqrt{z}}dz \end{aligned}$$

となりますから、$(X-m)^2$ の密度は

$$r(x) = \begin{cases} 0 & x < 0 \\ \frac{h(m+\sqrt{x})+h(m-\sqrt{x})}{2\sqrt{x}} & x \ge 0 \end{cases}$$

です。これを使って期待値を計算してみると

$$\begin{aligned} E[(X-m)^2] &= \int_{-\infty}^{\infty} xr(x)dx \\ &= \int_0^{\infty} x\frac{h(m+\sqrt{x}) + h(m-\sqrt{x})}{2\sqrt{x}}dx \\ &= \int_0^{\infty} x\frac{h(m+\sqrt{x})}{2\sqrt{x}}dx + \int_0^{\infty} x\frac{h(m-\sqrt{x})}{2\sqrt{x}}dx \end{aligned}$$

であり、第 1 の積分で $m + \sqrt{x} = z$、第 2 の積分で $m - \sqrt{x} = z$ とおけば

$$\begin{aligned} &= \int_m^{\infty}(z-m)^2h(z)dz + \int_m^{-\infty}(z-m)^2h(z)(-1)dz \\ &= \int_{-\infty}^{\infty}(z-m)^2h(z)dz \end{aligned}$$

です。

2.2.3　分散

データの各数値が平均値から（平均的に見て）どの程度ずれているかを調べようとして、単純に各数値と平均値との差をとって $X - E[X]$ を考えてもその平均値は

$$E[X - E[X]] = E[X] - E[X] = 0$$

となってしまいます。つまり『ずれ』そのものの平均値はずれが１の部分と－１の部分が打ち消し合って０になってしまい埒があきません。そこで右にずれるのも左にずれるのも区別しない事にしてずれの絶対値をとって $E\left[|X-E[X]|\right]$ とするのですが、絶対値を含む計算は面倒になりますからこれもイマイチです。そこで代替物として自乗 $E\left[(X-E[X])^2\right]$ を使う事にします。

定義 2.2.1 データ X が平均値 $E[X]$ をもち、更に $E\left[(X-E[X])^2\right]$ も存在する場合これをデータ X の分散（variance）と言って記号 $Var[X]$ で表します。X の密度関数が $h(x)$ であればこれは

$$Var[X]=E\left[(X-E[X])^2\right]=\int_{-\infty}^{\infty}(x-E[X])^2h(x)dx$$

です。また、分散の正の平方根を標準偏差（standard deviation）と言います。

分散は差の自乗ですから例えば長さのデータなら単位がメートルの自乗になってしまっています。そこでスケールを元に戻した標準偏差を使う事も大いにあります。

また、密度関数 $h(x)$ をもつデータの場合に分散を計算してみると

$$\begin{aligned}Var[X]&=E\left[(X-E[X])^2\right]\\&=\int_{-\infty}^{\infty}(x-E[X])^2h(x)dx\\&=\int_{-\infty}^{\infty}x^2h(x)dx-2E[X]\int_{-\infty}^{\infty}xh(x)dx+E[X]^2\int_{-\infty}^{\infty}h(x)dx\\&=E\left[X^2\right]-2E[X]^2+E[X]^2\\&=E\left[X^2\right]-E[X]^2\end{aligned}$$

となっている事が分かります（同様の事は高々可算無限データに対しても成立します）。場合によってはこの形の方が便利な場合もありますので覚えておくべきです。

事実 2.2.2 一般に、密度関数 $h(x)$ をもつデータ X と関数 $F(x)$ に対して積分 $\int_{-\infty}^{\infty}F(x)h(x)dx$ が有限値として存在するとき、この積分は $E[F(X)]$ に等しくなります。

2.3 問題演習

基本演習 2.1 次のデータ X について、$E[X]$、$E[X-3]$、$E\left[(X-3)^2\right]$ を計算して下さい：

数値	7	6	5	4	3	2	1	0
相対度数	$\frac{2}{12}$	$\frac{1}{12}$	0	0	$\frac{3}{12}$	$\frac{2}{12}$	$\frac{3}{12}$	$\frac{1}{12}$

基本演習 2.2 X の密度関数が次の $f(x)$ で与えられているとき定数 a の値を求め、平均値 $E[X]$、分散 $Var[X]$ を求めて下さい。

$$f(x)=\begin{cases}a(2-x) & 0\leq x\leq 2\\ 0 & \text{otherwise.}\end{cases}$$

基本演習 2.3 区間 $[-1,2]$ 上の一様分布に従う確率変数 X、つまり、密度関数が次の $h(x)$ であるような確率変数 X：

$$h(x)=\begin{cases}\frac{1}{3} & -1\leq x\leq 2\\ 0 & \text{otherwise}\end{cases}$$

について、確率 $P[1<X<5]$ と期待値 $E[X]$、分散 $Var[X]$ を求めて下さい。

基本演習 2.4 有限データの場合にも $Var[X]=E\left[X^2\right]-E[X]^2$ を証明して下さい。

基本演習 2.5 密度 $f(x)$ をもったデータ X の平均値が m、分散が v であるとき、派生データ $-2X+3$ の平均値、分散を m,v で表して下さい。

発展演習 2.6 X の密度関数は $h(x)$ で分散が存在するとし、任意の $\alpha>0$ を１つとります。分散の計算式の積分：

$$Var[X]=\int_{-\infty}^{\infty}(x-E[X])^2h(x)dx$$

において、積分範囲を $|x-E[X]|\geq\alpha$ である x の範囲と $|x-E[X]|<\alpha$ である x の範囲に分けて後者を捨てることによって Chebyshev の不等式：

$$P[|X-E[X]|\geq\alpha]\leq\frac{1}{\alpha^2}Var[X]$$

が成立する事を示して下さい。またこの不等式からどんな事が読み取れますか。

問題演習解答

基本演習 2.1

$$E[X] = 7 \cdot \frac{2}{12} + \cdots + 0 \cdot \frac{1}{12} = 3$$
$$E[X-3] = (7-3)\frac{2}{12} + \cdots + (0-3)\frac{1}{12} = 0$$
$$E[(X-3)^2] = (7-3)^2\frac{2}{12} + \cdots + (0-3)^2\frac{1}{12} = \frac{16}{3}$$

□

基本演習 2.2

$P[-\infty < X < \infty] = 1$ によれば

$$1 = \int_{-\infty}^{\infty} f(x)dx = \int_0^2 a(2-x)dx = \left[2ax - \frac{a}{2}x^2\right]_0^2 = 2a$$

となって $a = \frac{1}{2}$ が分かります。これを使って計算すれば

$$\int_{-\infty}^{\infty} xf(x)dx = \int_0^2 \frac{1}{2}x(2-x)dx = \frac{2}{3} = E[X]$$

であり

$$\int_{-\infty}^{\infty} x^2 f(x)dx - E[X]^2 = \int_0^2 x^2\frac{1}{2}(2-x)dx - \frac{4}{9} = \frac{2}{9} = Var[X]$$

です。 □

基本演習 2.3

$$P[1 < X < 5] = \int_1^5 h(x)dx = \int_1^2 \frac{1}{3}dx = \frac{1}{3}$$
$$\int_{-\infty}^{\infty} xh(x)dx = \int_{-1}^2 \frac{1}{3}x\,dx = \left[\frac{1}{6}x^2\right]_{-1}^2 = \frac{1}{2} = E[X]$$
$$\int_{-\infty}^{\infty} \left(x-\frac{1}{2}\right)^2 h(x)dx = \int_{-1}^2 \frac{1}{3}\left(x-\frac{1}{2}\right)^2 dx = \left[\frac{1}{9}\left(x-\frac{1}{2}\right)^3\right]_{-1}^2 = \frac{3}{4} = Var[X]$$

□

基本演習 2.4

有限データ X：$d_1, \ldots, d_n$ を考えます。平均値を m と書くことにすれば、

$$Var[X] = E[(X-m)^2] = \frac{1}{n}\sum_{j=1}^n (d_j - m)^2 = \frac{1}{n}\sum_{j=1}^n (d_j^2 - 2md_j + m^2)$$

であり、和を 3 つに分けて計算すると

$$Var[X] = \frac{1}{n}\sum_{j=1}^n d_j^2 - 2m\frac{1}{n}\sum_{j=1}^n d_j + m^2 = E[X^2] - m^2$$

となって確かに成り立ちます。 □

基本演習 2.5

$$E[-2X+3] = \int_{-\infty}^{\infty} (-2x+3)f(x)dx = -2\int_{-\infty}^{\infty} xf(x)dx + 3 = -2m + 3,$$

$$\begin{aligned} Var[-2X+3] &= E[(-2X+3-E[-2X+3])^2] \\ &= E[(-2X+3+2m-3)^2] \\ &= E[4(X-m)^2] \\ &= 4v. \end{aligned}$$

□

発展演習 2.6

X の密度関数が $h(x)$ で分散が存在するとき、任意の正の数 α に対して

$$\begin{aligned} Var[X] &= \int_{-\infty}^{\infty} (x-E[X])^2 h(x)dx \\ &= \int_{-\infty}^{E[X]-\alpha} (x-E[X])^2 h(x)dx \\ &\qquad + \int_{E[X]-\alpha}^{E[X]+\alpha} (x-E[X])^2 h(x)dx + \int_{E[X]+\alpha}^{\infty} (x-E[X])^2 h(x)dx \end{aligned}$$

と分けて右辺第 2 項を 0 で置き換えれば

$$\begin{aligned} &\geq \int_{-\infty}^{E[X]-\alpha} (x-E[X])^2 h(x)dx + \int_{E[X]+\alpha}^{\infty} (x-E[X])^2 h(x)dx \\ &\geq \int_{-\infty}^{E[X]-\alpha} \alpha^2 h(x)dx + \int_{E[X]+\alpha}^{\infty} \alpha^2 h(x)dx \\ &= \alpha^2 P[|X-E[X]| \geq \alpha] \end{aligned}$$

と評価されますから結果的に

$$P[|X-E[X]| \geq \alpha] \leq \frac{1}{\alpha^2}Var[X]$$

が（任意の $\alpha > 0$ に対して）成り立つ事が分かります。
分散が小さければ平均値からのずれが大きい確率は小さい事が分かります。 □