基礎・経済統計　6

確率分布

1．確率変数

- 事象を数値化したもの
  - （事象ー＞数値）の関数
- 自然に数値されている場合
  - さいころの目
  - 量的尺度
- 数値化が必要な場合
  - 質的尺度，順序的尺度
  - それらの尺度に数値を割り当てる
  - 例えば，コインの表が出たら1，裏なら0

2．離散確率変数と連続確率変数

- 確率変数の値
  - 連続値をとるもの
    - 身長，体重，実質GDPなど
  - とびとびの値＝離散値をとるもの
    - 新生児の性別：　男に1という値を割り振り，女には0を割り振る
      ＜質的尺度の数値化＞
    - ある夫婦の子供の数：0，1，2，3，4，…＜元々離散＞
- これらが確率的に決まる場合
  - 連続値なら連続確率変数
  - 離散値なら離散確率変数

度数分布表の応用による確率の表現

- 度数分布表
  - データ{4，3，6，5，3，6，4，1，2，5，1，2}

階級	階級値	相対度数	累積相対度数
0．5～1．5	1	1／6	1／6
1．5～2．5	2	1／6	2／6
2．5～3．5	3	1／6	3／6
3．5～4．5	4	1／6	4／6
4．5～5．5	5	1／6	5／6
5．5～6．5	6	1／6	6／6

- さいころの確率の「確率分布表」

区間	確率変数値	確率	累積確率
0．5～1．5	1	1／6	1／6
1．5～2．5	2	1／6	2／6
2．5～3．5	3	1／6	3／6
3．5～4．5	4	1／6	4／6
4．5～5．5	5	1／6	5／6
5．5～6．5	6	1／6	6／6

度数分布表，ヒストグラムを応用した確率のグラフ表現

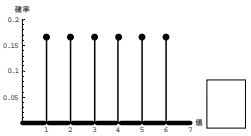
- 階級は区間と対応
- 相対度数と確率を対応させて考える
- 累積相対度数を累積確率に対応させる
- この対応をつかって確率に対する「度数分布表」，「累積度数分布表」，「ヒストグラム」，「累積ヒストグラム」などを考える
- それぞれの呼び方は確率を前につけて，「確率分布表」，「累積確率分布表」，「確率ヒストグラム」などとよぼう．
- また，この「確率分布表」を元にその平均，分散を求める．→確率分布に対する平均，分散
  - データの度数分布表を元に計算する平均，分散は標本平均，標本分散と以後呼ぶ

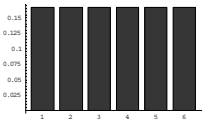
3．確率関数
3．1　概念

- 離散確率変数に限定
  - 離散確率変数の分布を特定する方法は？
- 飛び飛びの値それぞれになる確率を示す
  - 確率関数
    - $X = v_i \quad i = 1, \cdots, k$ ，つまり，確率変数Xはk個の飛び飛びの値をとるとする．
    - $p(t) = p_X(t) = \begin{cases} t = v_i \text{の場合} & P(X = v_i) \\ \text{それ以外の場合} & 0 \end{cases}$
- 確率関数を表にすると「確率分布表」になる．

３．２
確率関数とヒストグラム

- 確率関数のグラフ
  - ある値をとる確率（全事象のうちある値をとる割合）
- 確率関数はある意味で「確率ヒストグラム」の極限
  - ヒストグラムの縦軸として相対度数（全標本のうちある階級に属する割合）ではなく確率をとり，階級幅をどんどん縮めると確率関数のグラフが得られる．
- 例はサイコロの目





7

３．２の詳しい説明

- ヒストグラムの確率版
  - 確率変数値がある区間 $(a,b]$（階級）に属する確率をもとにヒストグラムを書く
  - $a<t\le b$ に対する縦軸の値は，$P[a<X\le b]$ となる．
  - a,bの間隔をどんどん狭めていく．
    - さいころの場合は $P[a<X\le b]$ は $P(X=t)$ に近づく
    - $t=1,2,3,4,5,6$ に関しては1/6，それ以外は0となる．
    - つまり，確率関数のグラフになる．

8

４．確率分布関数
４．１　概念

- 離散確率変数でも連続確率変数でも定義可
- 確率変数Xの分布関数　$F(t)=F_X(t)=P(X\le t)$
- 離散確率変数の場合
  - t 以下の値をとる確率の合計

$$F(t)=\sum_{i=1}^{k}p(v_i)\times I_{v_i\le t}(t)=\sum_{\substack{i=1\\ v_i\le t}}^{k}p(v_i)$$

$$I_{v_i\le t}(t)=\begin{cases}v_i\le t\text{の場合} & 1\\ \text{それ以外} & 0\end{cases}$$

9

４．２．
確率分布関数と「累積確率分布表」

- 確率分布関数はある値t以下の確率の合計
- 確率分布関数の表を作成すると「累積確率分布表」ができる．
- 「累積確率分布表」から「累積確率ヒストグラム」を作成する．
  - それは，確率分布関数のグラフとは違う，
  - 区間幅を狭めることによって，「累積確率ヒストグラム」を確率分布関数にいかようにも近づけうる．

10

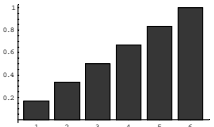
４．２の説明

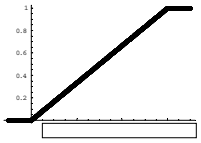
- 累積相対度数分布の確率版
  - 確率変数値がある区間 $(a,b]$（階級）に属する確率をもとに相対度数分布を書き，それを累積することで累積相対度数分布を求める．
  - $a<t\le b$ に対する縦軸の値は，$P(X\le b)$ となる．
  - a,bの間隔をどんどん狭めていく．
    - bがtに近づいていく
    - $P(X\le t)$ が縦軸の値になる
- 累積相対度数の確率版の極限が分布関数

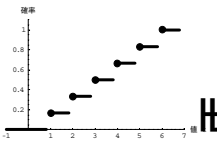
11

４．２の説明グラフ

- サイコロの目の累積相対度数グラフ（下は累積度数多角形）と分布関数グラフ







12

4．3　確率分布関数と区間確率

- 累積相対度数分布からある階級の相対度数を求める
  - ある階級の累積相対度数ーその直前の階級の累積相対度数
- 類推

  $P(a < X \le b) = P(X \le b) - P(X \le a) = F(b) - F(a)$
  - つまり，ある区間の確率<ある階級の相対度数>は，分布関数の区間の上限の時の値<その階級の累積相対>ー区間下限の時の値<その直前の階級の累積相対>できまる．

13

5．確率密度関数
5．1　連続確率変数と確率ヒストグラム

- 離散確率変数についてはヒストグラムに対応するものとして，確率関数のグラフが考えられた
- 連続確率変数とは確率分布関数が連続のもの
- 連続確率変数ではどうだろうか？
  - 離散の場合と同様にやってみると $P(X = t)$ を得る
  - しかし，$P(X = t)$ は連続確率変数の場合は0
  - つまり，連続確率変数の場合は，確率関数は0の値しかとらない．その意味でヒストグラムの極限は横軸に一致する．ー＞困った！

14

連続確率変数の場合 $P(X = t) = 0$ になる

$F(t) = \Pr[X \le t]$　はtに関して連続である．

それを利用するために，正の小さい数 ε に対して $F(t-\varepsilon) = \Pr[X \le t-\varepsilon]$ を考える．

$$F(t) - F(t-\varepsilon) = \Pr[X \le t] - \Pr[X \le t-\varepsilon]$$
$$= \Pr[t-\varepsilon < X \le t] \ge \Pr[X = t] \ge 0$$

となる．ところが，F（t）の連続性から，$\lim_{\varepsilon \downarrow 0}\{F(t) - F(t-\varepsilon)\} = 0$ となるので，上の式の最左辺も $\lim_{\varepsilon \downarrow 0} \Pr[X = t] = \Pr[X = t] = 0$

15

5．2　連続確率変数と修正ヒストグラム

- ヒストグラムの場合，棒グラフの面積の合計は1ではない．
- ヒストグラムの面積の合計が1になるようにしよう．
  - （階級幅×棒の高さ）の合計＝1になるようにする
  - 相対度数の合計=1
  - 棒の高さ＝相対度数／階級幅にすればよい．
  - そうすれば，ある階級の累積相対度数は，その階級までのヒストグラムの棒の面積の合計
- 修正ヒストグラムと呼ぼう

16

5．2　連続確率変数と修正ヒストグラム

- 連続確率変数の場合
  - 修正ヒストグラムの確率版

    tが $a < t \le b$ のときの，縦軸＝棒の高さ $\dfrac{P(a < X \le b)}{b-a}$
  - 幅を0に近づけたときの極限
    - 修正ヒストグラムの極限グラフ
    - つまり，横軸がtのとき，縦軸が $\lim\limits_{\substack{a \to t \\ b \to t}} \dfrac{P(a < X \le b)}{b-a}$
- これを確率密度関数のグラフと呼ぶ

17

5．3　確率密度関数の概念

- 確率密度関数

$$f(t) = f_X(t) = \lim_{\substack{a \to t \\ b \to t}} \frac{P(a < X \le b)}{b-a} = \lim_{\substack{a \to t \\ b \to t}} \frac{F(b) - F(a)}{b-a} = F'(t)$$

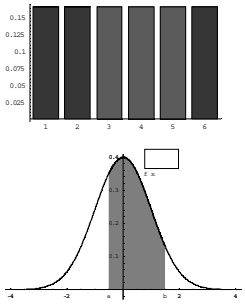
  - なぜ密度か？

    $\dfrac{P(a < X \le b)}{b-a}$ は確率を区間の長さ＝1次元面積で割っているので確率の密度と考えられる．
    - a,bをtに近づけているのでtという点での確率密度

18

５．４　確率と確率密度関数（１）

- 修正ヒストグラムの棒の面積のある階級まで合計と，一つ前の階級までの棒の面積の合計の差がある階級の相対度数（確率）
- この考え方を修正ヒストグラムの極限である確率密度関数に適用しよう―＞右図の灰色の面積が $P(a < X \le b)$



19

５．４　確率と確率密度関数（２）

- 灰色の面積は密度関数の定積分で表せるから，
$$P(a < X \le b) = P(a \le X \le b) = \int_a^b f(t)dt$$
- また $a \to -\infty$ とすると
$$P(X \le b) = \int_{-\infty}^b f(t)dt = F(b)$$
- では，離散確率変数に確率密度関数はあるか？
  - 離散の場合，ヒストグラムの極限は確率関数
  - 確率密度関数は修正ヒストグラムの極限
  - 離散の場合，修正ヒストグラムは，確率/階級幅
  - ありえる値のところでは，確率関数が正の値をとるので，階級幅を0に近づけると，修正ヒストグラムの極限＝確率密度は無限大
  - よって，離散の場合，確率密度関数は存在しない．

20

６．５
確率関数と密度関数の基本性質

- 確率関数の場合
  - 確率によるヒストグラムの極限だから関数値＝ヒストグラムの高さの合計は確率の合計＝１
  - 確率関数の合計は１　$\sum_{i=1}^{k} f(v_i) = 1$
- 確率密度関数の場合
  - ヒストグラムの面積が1になるようにした修正ヒストグラムの極限だから，密度関数の面積合計も1

$$\int_{-\infty}^{\infty} f(t)dt = F(\infty) - F(-\infty) = P(X \le \infty) - P(X \le -\infty)$$
$$= 1 - 0 = 1$$

21

６．分布の代表値

- データの場合のアナロジー
  - 相対度数によるヒストグラム→代表値
  - 確率分布によるヒストグラム→代表値
- 分布の代表値
  - 分布の平均（母平均）または期待値
  - 分布の分散（母分散）
  - 分布のパーセント点

22

６．１　分布の平均（母平均，期待値）

- 分布の重心
- 計算法
  - 離散確率変数の場合
    - とりうる値に対してその値になる確率（その値に対する確率関数の値）をかけたものの合計

$$E[X] = \mu = \mu_X = \sum_{i=1}^{k} v_i P(X = v_i) = \sum_{i=1}^{k} v_i\, p(v_i)$$

  - 連続確率変数の場合

$$E[X] = \mu = \mu_X = \int_{-\infty}^{\infty} t\, f(t)dt$$

23

６．２　分布の分散（母分散）

- 確率分布の散らばりの指標
- 計算法
  - 離散確率変数の場合

$$V[X] = \sigma_X^2 = \sum_{i=1}^{k} (v_i - \mu_X)^2\, p(v_i)$$

  - 連続確率変数の場合

$$V[X] = \sigma_X^2 = \int_{-\infty}^{\infty} (t - \mu_X)^2\, f(t)dt$$

24

6．3
確率変数から新たな確率変数を作る

- 確率変数Xの関数もまた確率変数
  確率変数 $g(X)$ ができる
  たとえば，$3X^2 - X + e^X$. X=1となったときの
  この確率変数 $3X^2 - X + e^X$ の値は $31^2 - 1 + e^1 = e + 2$
- この確率変数 $g(X)$ の分布関数は，
  $$F_{g(X)}(t) = P(g(X) \le t) = P(X \le g^{-1}(t)) = F_X(g^{-1}(t))$$
- 期待値計算(viをg(vi)，tをに置き換える)

$$E[g(X)] = \sum_{i=1}^{k} g(v_i)p(v_i) \qquad E[g(X)] = \int_{-\infty}^{\infty} g(t)f(t)dt$$

25

6．3 期待値，分散の演算（1）

- 期待値の性質
  - 離散の場合

$$E[X - \mu_X] = \sum_{i=1}^{k}(v_i - \mu_X)p(v_i)$$
$$= \sum_{i=1}^{k} v_i p(v_i) - \mu_X \sum_{i=1}^{k} p(v_i) = \mu_X - \mu_X = 0$$

  - 連続の場合

$$E[X - \mu_X] = \int_{-\infty}^{\infty}(t - \mu_X)f(t)dt$$
$$= \int_{-\infty}^{\infty} t\, f(t)dt - \mu_X \int_{-\infty}^{\infty} f(t)dt = \mu_X - \mu_X = 0$$

26

6．3 期待値，分散の演算（2）

- 期待値の演算
  - X,Yは確率変数，a,bは確率変動しないとする

$$E[aX + bY] = aE[X] + bE[Y]$$

- 分散の演算

$$V(aX + b) = a^2 V(X)$$

  - XとYが独立の場合

$$V(aX + bY) = a^2 V(X) + b^2 V(Y)$$

27

6．4 分布のパーセント点

- 確率変数Xの分布の α %点
  $F(t) = P(X \le t) = \alpha/100$ となるtの値
- 分布の中央値（メジアン）
  $F(t) = P(X \le t) = 0.50$ となるtの値

28

7．正規確率変数と正規分布
7．1 独立な変数の和の分布（1）

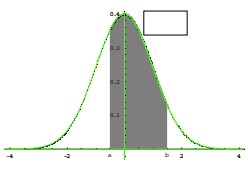
- 独立な確率変数の和の分布を考える
  - $X_1, X_2, \cdots, X_n$ を独立で期待値 $E[X_i] = 0$，分散 $V(X_i) = 1$の確率変数の列とする
  - 例えば，コインを繰り返し投げる場合，i回目に投げたときに表がでると1，裏がでると-1の値をとるような確率変数を X_i とする．この場合，平均0で分散が1の確率変数列になる
- このとき，$S_n = X_1 + X_2 + \cdots + X_n$ は

$$E[S_n] = E(X_1) + \cdots + E(X_n) = 0 + \cdots + 0$$
$$V(S_n) = V(X_1) + \cdots + V(X_n) = \underbrace{1 + \cdots + 1}_{n\text{個}} = n$$

29

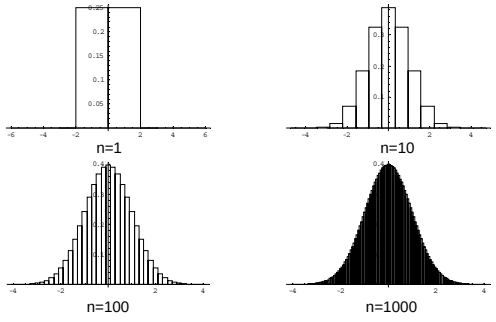
7．1 独立な変数の和の分布（2）

- Snをその標準偏差 $\sqrt{n}$ で割る $V[S_n/\sqrt{n}] = V(S_n)/(\sqrt{n})^2 = n/n = 1$
  - 一般的にある確率変数をその標準偏差で割って得られる確率変数は分散，標準偏差ともに1．
  - Snは平均0，分散1の確率変数．
- さらにnを大きくしていくと $S_n/\sqrt{n}$ の密度関数はきれいな釣り鐘型をする．



30

Snの修正確率ヒストグラムの推移



31

「正規確率変数に近づく」について（1）

- 修正ヒストグラムでの階級の決め方
  - Snの値はnが奇数の場合奇数，nが偶数の場合は偶数になる．従って，Snの値の間隔は2．取りうる値同士の真ん中に階級の境目を持ってくる．

$$P(S_n = k) = P(k-1 < S_n \le k+1)$$

$$P\left(\frac{S_n}{\sqrt{n}} = \frac{k}{\sqrt{n}}\right) = P\left(\frac{k-1}{\sqrt{n}} < \frac{S_n}{\sqrt{n}} \le \frac{k+1}{\sqrt{n}}\right)$$

  - 連続補正の根拠

32

「正規確率変数に近づく」について（2）

- 別の階級の決め方では？

$$P(S_n = k) = P(k-1/2 < S_n \le k+1/2)$$

$$P\left(\frac{S_n}{\sqrt{n}} = \frac{k}{\sqrt{n}}\right) = P\left(\frac{k-1/2}{\sqrt{n}} < \frac{S_n}{\sqrt{n}} \le \frac{k+1/2}{\sqrt{n}}\right)$$

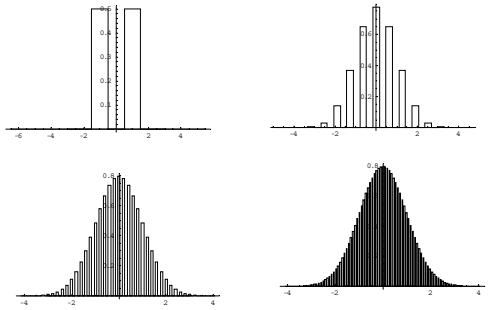
  - つまり階級幅を半分で考える．
  - この場合は，

$$P(S_3 = 0) = P(-1/2 < S_n \le 1/2) = 0$$

    - つまり，修正ヒストグラムは，そこでは0．

33

Snの修正確率ヒストグラムの推移 連続補正しない場合



34

「正規確率変数に近づく」について（3）

- 連続補正に対応しない階級幅の取り方をすると，修正ヒストグラムは極限は連続な密度関数にならない．
  - 連続確率分布での近似は出来ない．
  - 今回は，$S_n/\sqrt{n}$ の離散確率分布がnがどんどん大きくなるにつれて連続確率分布に近づくことを示したいので，このような修正ヒストグラムではその様子はわからない．
- 同時になぜ連続補正が必要かも示している．

35

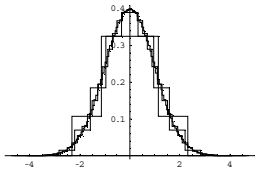
7．1 独立な変数の和の分布（3）

- $S_n/\sqrt{n}$ の極限分布
  - 標準正規分布とよぶ（N(0,1)と書く）
  - 密度関数 $f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$
  - このような分布を持つ確率変数をZとする．
- 一般の正規分布
  - 平均 μ，分散 σ^2（標準偏差 σ）の正規分布（$N(\mu, \sigma^2)$）は確率変数 $\sigma Z + \mu$ の分布

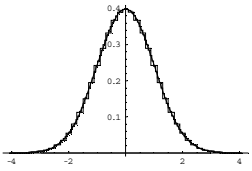
$$f(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$

36

修正確率ヒストグラムの極限が例の密度関数になっている



n=3,10,100,1000について修正ヒストグラムの上部のみを重ねて描いた



n＝100,1000の修正ヒストグラムと前ページの関数を重ねて書いた

7．2 正規分布表

- 逆に平均μ，標準偏差σの正規分布は標準正規分布の確率変数Zを使って$X=\sigma Z+\mu$で表せるから，$\frac{X-\mu}{\sigma}=Z$ はN(0,1)の分布となる．
- 従って，

$$P(X\le s)=P\left(\frac{X-\mu}{\sigma}\le\frac{s-\mu}{\sigma}\right)=P\left(Z\le\frac{s-\mu}{\sigma}\right)$$

となり，
確率は標準正規分布の確率変数が $\frac{s-\mu}{\sigma}$ 以下になる確率

- 教科書p．279の標準正規分布表を使えば計算可

注意

- 正規分布は平均の値を軸にして左右対称な密度関数を持つ
- 故に，平均の値を軸にして左右対称な確率分布関数を持つ
  - よって，$P(Z\le a)=P(-Z\le a)=P(Z\ge -a)$
  - または，$P(Z\le -a)=P(-Z\ge a)=P(Z\ge a)$
- 例えば，

$$P(Z\le -1)=P(-Z\ge 1)=P(Z\ge 1)=1-P(Z<1)$$
$$=1-P(Z\le 1)$$

7．3 偏差値の意味

- 正規分布している変数の場合，偏差値が特定の値以下である確率は標準正規分布表から求められる．
  - 偏差値S $S=\frac{X-\mu}{\sigma}\times 10+50$
  - 仮定の下では $S=Z\times 10+50$
  - よって偏差値は仮定の下で平均50，標準偏差10の正規分布
  - 従って，

$$P(S\le t)=P\left(\frac{S-50}{10}\le\frac{t-50}{10}\right)=P\left(Z\le\frac{t-50}{10}\right)$$

例

- ある変数が正規分布に従っているとして，その偏差値が65以下になる確率は，

$$P\left(Z\le\frac{65-50}{10}\right)=P(Z\le 1.5)=0.9332$$

- 逆にいうと偏差値65を上回る確率は6.68%
- まあ，一人一人の試験の点をみたときそれが，正規分布に従う確率変数であることは少ないですが．．．