第3回 確率変数と確率分布

2007年9月25日

前回のおさらい

- 度数分布表，ヒストグラムの作成
  - 分析ツールの「ヒストグラム」を使った
- 平均・分散などの統計量の計算
  - Excelの関数を使った
  - 分析ツールの「基本統計量」からも計算できる

今日のお題

- 確率変数
- 確率分布

- ちなみに・・・

  今期の目標は2種類のデータが与えられた際に，そのデータに差があるかどうかを調べる（「検定」と呼ぶ）方法を習得すること
  - そのためには，確率分布のことについて多少なりとも知っておく必要がある
  - 今日はその準備として確率変数・確率分布について軽く勉強します

確率変数・確率分布

- 確率変数 X
  - 変化する量 X のとる値に対して確率が定まるもの
  - 離散型確率変数
    - 確率変数が有限個（もしくは可算個）の値をとる時
      - サイコロの出た目
  - 連続型確率変数
    - 確率変数の値が連続的に変化するとき
      - ある集団における身長や体重
- 確率分布
  - 確率変数がどのような値になるかという法則性を与えるもの

確率分布の平均・分散・標準偏差

- 確率分布の分布関数を $f(x_k)$
- 平均（期待値） $E(X)=\mu=\begin{cases}\sum_{k=1}^{n} x_k f(x_k) \\ \int_{-\infty}^{\infty} x f(x)dx\end{cases}$
- 分散

$$\mathrm{V}(X)=E\left[(X-\mu)^2\right]=\begin{cases}\sum_{k=1}^{n}(x_k-\mu)^2 f(x_k) \\ \int_{-\infty}^{\infty}(x_k-\mu)^2 f(x)dx\end{cases}$$

- 標準偏差 $\sigma(X)=\sqrt{\mathrm{V}(X)}$

共分散

- 2つの確率変数 X,Y の平均を μ_x,μ_y
- $X-\mu_x, Y-\mu_y$ の積を確率変数とする確率分布の平均が共分散
- 共分散 $\mathrm{Cov}(X,Y)=E\left[(X-\mu_x)(Y-\mu_y)\right]$

- 相関係数：共分散を標準偏差の積で割った値

$$\rho(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)}$$

平均・分散のいくつかの性質

$E(aX^2+bX+c)=aE(X^2)+bE(X)+c$

$E(X+Y)=E(X)+E(Y)$

$E(XY)=E(X)E(Y)$　X,Yが独立の時

$V(X)=E(X^2)-E(X)^2$

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

$Cov(X,Y)=E(XY)-\mu_x\mu_y$

$Cov(X,Y)=0$　X,Yが独立の時

$V(X+Y)=V(X)+V(Y)+Cov(X,Y)$

$V(X+Y)=V(X)+V(Y)$　X,Yが独立の時

難しい話は置いといて

- 分布の形を見てみよう
  - 立教大学経営学部経営学科の山口先生のホームページにあるツールを使わせてもらう
    - http://www.ir.rikkyo.ac.jp/~kazunori/stat/distribution/index.html
- 実際に分布について知るためにExcelで遊んでみよう
- 今後のスライドには分布の定義式とかを書いてあるが，基本的にはそれぞれの分布がどういう形なのか知ってもらえれば良い

離散型1

- 離散一様分布

$UD(n;a,d),\ -\infty<a<\infty,\ d>0,\ f(a+kd)=1/n,\ k=0,1,\ldots,n-1$

  - 例：サイコロ
  - 分析ツール→乱数発生→離散
    - 変数の数：乱数を発生させたい列の数
    - 乱数の数：乱数を発生させたい行の数
    - 値と確率の入力範囲

離散型2

- 2項分布

$B(n;p),\ 0<p<1,\ f(k)={}_nC_k\,p^k(1-p)^{n-k},\ k=0,1,\ldots,n$

  - 例：くじ
  - 分析ツール→乱数発生→二項
    - 変数の数，乱数の数
    - p値：あたりの確率
    - 試行回数：くじを引く回数
  - 3項分布（多項分布）というのもある

離散型3

- ポアソン分布

$Po(\lambda),\ \lambda>0,\ f(k)=\frac{\lambda^k}{k!}e^{-\lambda},\ k=0,1,\ldots$

  - 例：窓口に到着する人の数の確率，機械が故障する確率
  - 分析ツール→乱数発生→ポワソン
    - 変数の数，乱数の数
    - k　：単位時間中に平均で事象が起こる回数
    - λ　：単位時間中に事象が起こる回数

連続型

- 正規分布

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

  - 例：身長・体重の誤差
  - 分析ツール→乱数発生→正規
    - 変数の数，乱数の数
    - 平均
    - 標準偏差

他にも

- 連続型でよく出てくる分布
  - t分布
  - カイ2乗分布
  - F分布
  - etc.

何で分布が必要なのか

- 検定をするときに使う
  - 検定をしたい値（たとえば平均）がある分布に従っていることが分かっているから
  - 有意か有意でないか（差があるかないか）の判断の時に分布の値に基づいた判断が行われる
  - 分布についてよくわからない場合・・・
    - そういうものだと思ってください
    - むしろ「検定」の方法・使い方・結果の読み取り方を知っておいたほうが役に立つ・・・はず