確率・統計の初歩(2)

別所俊一郎

いくつかの確率分布。

- 表現の約束
  - 確率変数 X の分布関数が F で表現されるとき、「確率変数 X は分布 F に従う」といい、「 $X \sim F$ 」と表す
  - 分布関数を大文字（ F ）、密度関数を小文字（ f ）で表す
  - 確率変数 X の密度関数を $f_X(x)$ と書くこともある
- よく使う確率分布
  - 検定で用いる
  - 正規分布（normal）、カイ2乗分布（Chi-squared）、
  - F分布、t分布
  - もちろん、他にもいろいろな名前の分布がある

正規分布normal distribution

- 釣鐘状の連続分布
- 数式で書くとやや複雑

$f_Y(y)$=

- 平均μ、分散σ^2とすると、区間 [μ -1.96σ, μ+1.96σ] に95%が含まれる
- 3次以上のモーメントが定数になり、平均と分散だけで分布が決まる
- 確率の分野では非常によく用いられる分布

正規分布normal distribution

- いくつかのお約束
  - 平均μ、分散σ^2の正規分布を $N(\mu, \sigma^2)$ と表す
  - 平均0、分散1の正規分布（ $N(0,1)$ ）のことをとくに「標準正規分布 standard normal」とよび Z で表す。すなわち、$Z \sim N(0,1)$
  - 標準正規分布の分布関数を Φ 、密度関数を ϕ で表す。すなわち定数 c に対して、$\Phi(c) = \Pr(Z < c)$
  - 正規分布に従う確率変数から、その平均を引いてその標準偏差で除して新しい確率変数を定義することを「標準化 normalized」と呼ぶ。たとえば、$Y \sim N(1, 4)$ のとき、

    $\Pr(Y < 2)$ =

多変量正規分布
multivariate normal

- いくつかの確率変数の同時分布に対して多変量正規分布を考えることができる。n変量正規分布は、平均の組合せ（ n 個）と、分散・共分散の組合せ（n^2個）によって定義できる。
- 確率変数 (X, Y) が2変数正規分布に従うとき、定数 a, b に対してその線形結合もまた正規分布に従い、

$$aX + bY \sim N$$

- 一般に、n変量正規分布に従う n 個の確率変数の線形結合もまた正規分布に従う

多変量正規分布
multivariate normal

- 一般に、n変量正規分布に従う n 個の確率変数について、各確率変数の周辺分布もまた正規分布となる
- n変量正規分布に従う n 個の確率変数のうち、共分散がゼロであるような2個の確率変数は互いに独立である
  - 互いに独立である確率変数の共分散は、同時分布の形状に関わらず、ゼロ。
  - 一般には、共分散がゼロであるからといって、2個の確率変数が独立であるとは限らない。
  - この性質は、正規分布では3次以上のモーメントが定数であることによる。

カイ2乗分布 χ^2, Chi-squared

- 仮説検定で用いられることの多い分布
- m個の独立した標準正規分布に従う確率変数の2乗和の分布を、自由度 m のカイ2乗分布とよぶ
  - 互いに独立な3つの確率変数が、$Z_1 \sim N(0,1)$, $Z_2 \sim N(0,1)$, $Z_3 \sim N(0,1)$ であるとき、
    $Z_1^2 + Z_2^2 + Z_2^2 \sim \chi^2(3)$
- 分布表についてはappendix参照

F分布、t分布

- 自由度 m のχ^2分布に従う確率変数を m で除して得られる確率変数の分布を自由度m のF分布とよぶ

- 標準正規分布に従う確率変数を、それとは独立に自由度 m のχ^2分布に従う確率変数を m で除したもので割って得られる確率変数の分布を、自由度 m の（スチューデント）t分布とよぶ
  - 自由度 m が十分に大きければ、t分布は正規分布に近い。極限では正規分布に一致する。$t_\infty = N(0, 1)$

無作為抽出 random sampling

- 母集団と標本
  - 検討の対象とする主体全体を母集団(population)とよび、そこから抽出された観測値(observation)の集合を標本(sample)とよぶ。
  - 母集団をすべて調査できるとき、悉皆調査とよぶ
  - 母集団として何を考えているかは重要
  - 手許にある標本から母集団の特性を推し量ることが計量経済分析の目的のひとつ
- 標本の抜き出し方は確率的
  - ほとんどのケースで、母集団から標本をどのように抽出(サンプリング)するかは確率的
  - 片寄りなくでたらめに標本を選ぶことを無作為抽出とよぶ
  - 無作為抽出では、ある抽出は、次に抽出されるものについての情報をまったくもたない。

9

標本統計量 sample statistics

- 標本の選び方が確率的であるから、標本の特性を表す指標(平均など)は確率変数である。
  - 無作為抽出の場合は、各観測値は互いに独立な確率変数であり、それらから計算される平均や分散もまた確率変数
  - 標本から計算される平均などを「標本統計量」という
  - 標本から計算される平均値は確率変数であるから、標本平均は確率分布を持つ。それゆえ、「平均の分散」を考えることができる。
  - 標本平均は母平均(母集団の平均)とは一致しないから、標本平均を「評価」するためには標本平均の分散や標準偏差を知る必要

10

i.i.d.

- 標本抽出を行うと、各観測値は確率変数だから、それらの多変量確率分布（同時分布）を考えることができる
  - 各観測値が同じ母集団から抽出され、各観測値の周辺分布が同じであるとき、それら観測値は同じ分布に従う identically distributed という
  - 無作為抽出などにより、各観測値の分布が互いに独立であるとき、それらは独立に分布している independently distributed という
  - 各観測値が独立に同じ分布に従う independently and identically distributed とき、略して i.i.d. である（独立同一分布）、という

標本平均 sample average (mean)

- 標本平均は、各観測値の平均で定義される

- 各観測値は確率変数だから、標本平均も確率変数。標本が変われば標本平均も変わる
- 標本平均の平均や、標本平均の分散を考えることができる
  - 抽出を繰り返し行い、そのたびごとに標本平均を計算したとき、その標本平均の平均や分散のこと。
  - 標本平均の分布は、標本の大きさにも依存

標本平均の分布

- 各観測値 (Y_1, Y_2, …, Y_n) がi.i.d.で、その平均と分散（母集団の平均と分散）を μ_Y, σ_Y^2 とする。
  - たとえば、観測値数が2のとき、標本平均の期待値は、
  - 一般に、標本平均の平均は、
  - すなわち、母集団の平均に一致する。

13

標本平均の分散

- 各観測値 (Y_1, Y_2, …, Y_n) がi.i.d.で、その平均と分散（母集団の平均と分散）を μ_Y, σ_Y^2 とする。
  - たとえば、観測値数が2のとき、標本平均の分散は、
  - 一般に、標本平均の分散は、
  - すなわち、標本が大きいほど小さい。

14