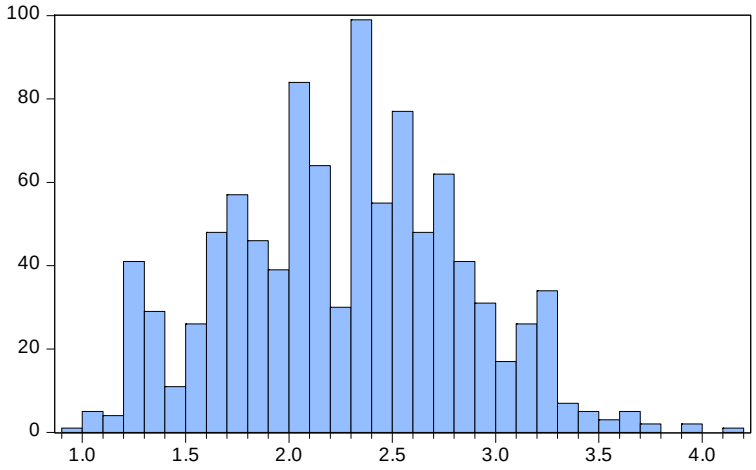
2014 年 2 月

データの分布を知る

前回はデータの平均や標準偏差を銘柄ごとにテーブルにまとめるプログラムを作成しました。今回はデータの分布を知るためのコマンドと、それに関連したプログラミングを例によって練習します。
ここで利用するデータは EViews に用意されている賃金データのファイル cps88.wf1 です。このファイルを開いて、シリーズ lwage で View/Descriptive Statistics and Tests/Histogram and Stats と操作して次のグラフを作成します。



Series: LWAGE
Sample 1 1000
Observations 1000

Mean	2.275496
Median	2.302585
Maximum	4.174387
Minimum	0.955511
Std. Dev.	0.563464
Skewness	0.077570
Kurtosis	2.640649
Jarque-Bera	6.383398
Probability	0.041102

「データの分布について知りたい」という場合、基本的には平均、分散、歪度、尖度を用いて分布の特徴付けを行うのが、統計学での基本的なアプローチです。
変数 lwage は、時給の対数をとった米国のデータで、データの個数は 1000 人分であることが分かります。平均は 2.27、中央値は 2.30、最大最小はそれぞれ 4.17,0.96 ほどになっています。Skewness(歪度)は 0.07、Kurtosis(尖度)は 2.64 です。正規分布を基準として考えたときに、歪度は横方向の偏りで、尖度は高さ方向のとがり具合です。ある変数が正規分布する場合、歪度は 0、尖度は 3 となります。
いま、歪度は 0.07、尖度は 2.64 ですが、このデータは「正規分布している」と言えるでしょうか。その検定に利用するのが Jarque-Bera 統計量です。今、p 値は約 4%となっています。帰無仮説は「このデータの分布は正規分布である」です。有意水準 5%とした場合、帰無仮説は棄却されます。つまり、正規分布しているとは主張できません。

さて、EViewsには様々な関数が予め用意されています。例えば、平均から尖度まで求めるためのコマンドを次に示します。

平均	@mean(変数名)
中央値	@median(変数名)
最大値	@max(変数名)
最小値	@min(変数名)
標準偏差	@stdev(変数名)
歪度	@skew(変数名)
尖度	@kurt(変数名)

例えば、lwageの平均値が知りたい場合はコマンドウィンドウに次のように入力します。

```
show @mean(lwage)
```

これらの関数コマンドはEViewsのオンラインヘルプでキーワード検索すれば、簡単に見つけることができます。

さて、EViewsの英語のPDFマニュアルや印刷された日本語マニュアルには、これら統計量の計算式が載っています。ここでは次に示す標準偏差、尖度、歪度、Jarque-Bera統計量を計算するプログラムを作成してみましょう。

標準偏差 $s = \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})/N - 1}$

歪度 $S = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \bar{y}}{\hat{\sigma}}\right)^3$

$\hat{\sigma} = s\sqrt{(N-1)/N}$

尖度 $K = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \bar{y}}{\hat{\sigma}}\right)^4$

$$Jarque - Bera = \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

先に計算した値を後段の統計量の計算に利用することが分かります。EViews のプログラミング例を次に示します。File/New/Program と操作して、次に示すプログラムを入力します。すべての入力が完了したら、最初にプログラムを「lwage.prg」という名前で保存します Run ボタンをクリックしてプログラムを実行します。

```
'1000個のデータがあります。
smpl @all
'個々の統計値を求めます。
'データlwageの個数を!nとします。
!n=@obssmpl

'lwageの偏差の二乗をx2とします。
series x2=(lwage-@mean(lwage))^2

'不偏分散による標準偏差を!sとします。
!s=@sqrt(@sum(x2)/(!n-1))

'歪度を計算するため標本標本分散を!bsとします。
!bs=!s*@sqrt((!n-1)/!n)

'歪度を!skwとします。
series x3=((lwage-@mean(lwage))/!bs)^3
!skw=(1/!n)*@sum(x3)

'尖度を!kとします。
series x4=((lwage-@mean(lwage))/!bs)^4
!k=(1/!n)*@sum(x4)

'ここまでの計算結果を利用して、Jarque-Bera統計値!jbを求めます。
!jb=(!n/6)*(!skw^2+(!k-3)^2/4)

'計算結果の一覧を表示
show !s    !skw !k !jb

'確認のための元のヒストグラムを表示
show lwage.hist
```

せっかくのここまで計算したので、最後に Jarque-Bera 統計値も求めてみましょう。マニュアルにもありますように、Jarque-Bera 統計量は自由度 2 のカイ二乗分布から求めることができます。実際、プログラムに次の 1 行を足して計算してみましょう。

```
show 1-@cchisq(!jb,2)
```

最後にでてきたカイ二乗分布は正規分布と並んで、計量分析ではよく利用される確率分布です。ここまで操作できた方は CPS88 を上書き保存して閉じます。

後半はカイ二乗分布について「馴染む」ために EViews を使ってシミュレーションをしてみましょう。カイ二乗分布の定義式を確認します。今、標準正規分布に従う、独立な確率変数 Zn を考えますと、次に示す Y は自由度 n のカイ二乗分布に従います。

$$Y = Z_1^2 + \cdots + Z_n^2 \sim \chi_n^2$$

標準正規分布の乱数機能を利用して、自由度 2 なので 2 つの独立な乱数を取得し、上式にしたがって Y を計算します。この計算を例えば、1 万回行ってヒストグラムを作成すると、95 パーセンタイル(ゼロに近いほうから数えて 9500 番目のデータの位置)が有意水準 5%の境界値にほぼ等しいと考えられます。

早速、新しいプログラムウィンドウに次のコマンドを入力し、Run ボタンで実行してみましょう。

```
'Unstructuredのワークファイルchi2testを作成します。
wfcreate(wf=chi2test) u 10000

'二乗和のシリーズyを作成します。
series y

'10000回、標準正規分布の乱数発生を行います。
for !i=1 to 10000

'行列オブジェクトm1に2つの変数を作成します。
    matrix m1=@mnrnd(2,1)
'1つ目の正規乱数を行列オブジェクトから取り出します。
    !z1=m1(1,1)
'2つ目の正規乱数を行列オブジェクトから取り出します。
    !z2=m1(2,1)
'カイ二乗値の定義にしたがって、!yを計算します。
    !y=!z1^2+!z2^2
'二乗和をシリーズyの1行目から順に入れて行きます。
```

```
    y(!i)=!y

next

'作成したシリーズyのヒストグラムを作成します。
show y.hist

'yの95パーセンタイルを求めます。
show @quantile(y,0.95)
```

このプログラムが終了したら、ヒストグラムとヒストグラムにおける 95 パーセンタイルの横軸の位置を画面に表示します。おそらく、横軸の位置は 0.6 位になるのではないでしょうか。

自由度 2 のカイ二乗分布で 95%点を正確に調べる場合は、次のようにします。

```
show @qchisq(0.95,2)
```

この結果、EViewsは6.0350777…という値を表示します。先ほどのプログラムでサンプル数を増やしていくほど、横軸の位置はこの値に近づくことになります。

今回は EViews プログラミング機能を利用して基本的な統計量の計算と、カイ二乗分布のシミュレーションを行ってみました。次回は、今回算出した「尖度」、「歪度」について振り返り、データの分布ということについて、「分布関数」や「密度関数」というキーワードを元に考え方を整理してみたいと思います。

株式会社　ライトストーン