[問題] 上に書いた、「時間順序に並んでいるが一定の間隔をもっていないデータ」の例を挙げてみよ。

[解答] Webページへのアクセスログ, POSデータ

また、時系列データには**連続時間(continuous-time)**の時系列データと**離散時間(discreate-time)**の時系列データの区別があるが、この講義では、離散時間の時系列データの解析についてのみを行うこととする。

1.2 時系列解析の困難さ

さて、これから時系列分析について講義するわけだが、なぜ殊更に時系列データについては、特別な方法で当たらなければならないのだろうか。基礎的な統計学の知識だけで、時系列データを解析することはなぜ困難なのか。その理由について簡単に述べる。結論から言えば、データを生成する確率分布が**i.i.d.(independent and identically distribution)**になっていないからということになるのだが、少し丁寧に説明をしてみることにしよう。時系列データのもつ特徴を以下に挙げる。

1. 自己相関
2. 標本系列の唯一性
3. 時間による変化

これらが、時系列データを時系列データたらしめている要因であり、解析の困難さの原因となっている。以下の小節ではこれを解説しよう。

1.2.1 自己相関

時系列データの大きな特徴の一つに、ある時点でのデータがその前の値に依存しているという性質が挙げられる。**クロスセクション(cross section)**と呼ばれる、ある一時点での複数の観測値を持つデータ(普段我々はこれを相手に統計解析をしている)では、それぞれの観測値は**独立(independent)**であると仮定して差し支えないことが多い。従って、同一の確率分布から、複数の値が独立に発生していると考えて良いわけだが、時系列データの場合はそうはいかない。このように、自らの過去の姿に影響を受けて系列が決まっていく性質を、**自己相関(autocorrelation)**と呼ぶ。

自己相関は時系列のモデルを考える際には、大変重要な概念であり、その大きさは、**自己相関係数(autocorrelation coefficients)** ρ_kと呼ばれる量で計られる。これは、ある時系列y_tとそのラグkの系列y_{t-k}のピアソンの積率相関係数を計算することにより得られる値である。

$$\rho_k = \frac{Cov(y_t, y_{t-k})}{sd(y_t)sd(y_{t-k})} \qquad k\text{次自己相関係数}$$

```
#NAを除去
np.omit <- as.ts(na.omit(nikkei$N225.Close))
str(np.omit)
```

```
##  Time-Series [1:9283] from 1 to 9283: 6503 6563 6614 6689 6737 ...
```

```
#ラグ1の自己相関係数
cor(np.omit[-1], np.omit[-9283])
```

```
## [1] 0.9993897
```

[問題] ラグ2の自己相関係数を求めてみよ。

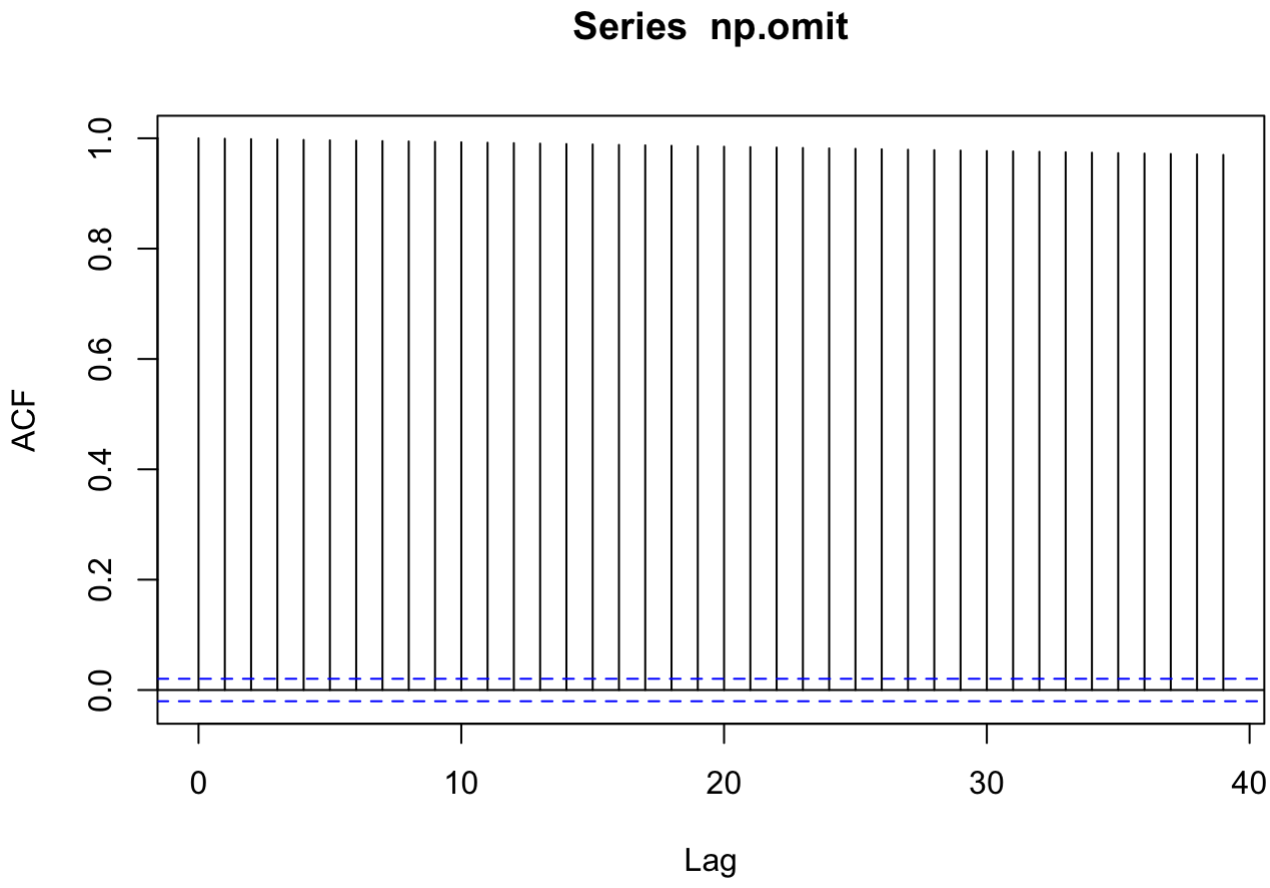
[解答] Rのscriptを以下に示す。

```
#ラグ2の自己相関係数
cor(np.omit[c(-1,-2)], np.omit[c(-9282,-9283)])
```

```
## [1] 0.9987802
```

さて、ここで自己相関はラグkの関数になっていることに注意しよう。このようにみた時、その関数を**自己相関関数(autocorrelation function)**といい、これをグラフ化したものを**コレログラム(correlogram)**と呼ぶ。

```
#コレログラム
acf(np.omit, plot=T)
```




```
#自己相関関数
acf(np.omit, plot=F)
```

```
##
## Autocorrelations of series 'np.omit', by lag
##
##    0    1    2    3    4    5    6    7    8    9   10   11
## 1.000 0.999 0.998 0.998 0.997 0.996 0.996 0.995 0.994 0.994 0.993 0.992
##   12   13   14   15   16   17   18   19   20   21   22   23
## 0.991 0.990 0.990 0.989 0.988 0.987 0.986 0.986 0.985 0.984 0.983 0.982
##   24   25   26   27   28   29   30   31   32   33   34   35
## 0.982 0.981 0.980 0.979 0.979 0.978 0.977 0.976 0.975 0.975 0.974 0.973
##   36   37   38   39
## 0.972 0.972 0.971 0.970
```

また、自己相関の有無を統計学的検定で調べることもできる。コレログラムにおける**青色の線**はこの検定の棄却域を示している。また1つ1つの自己相関ではなく、m次までに自己相関が少なくとも1つあることを検定することもできる。この検定は**Ljung-Box検定(Ljung-Box test)**と呼ばれていて、mを適切に選べば(これが難しいことがあるが)、検出力が高い検定として知られている。

1.2.2 標本系列の唯一性

統計学の基本は、**自分の持つデータがどのような確率分布から生成されたか**(このような確率分布を**母集団分布**もしくは**データ生成分布**という)を考えることである。これはすなわちデータが発生するメカニズムを**確率モデル(stocastic model)**として考えることともいえる。

さて、時系列のデータが発生するメカニズムを考えたい。例えば先ほどの例で、ある期における日経平均終値y_jが正規分布しているという仮定をおいたとしよう。

[問題] 上の仮定(ある期における日経平均終値が正規分布している)を置いたら、我々が次に求めなければならないことは何か。

[解答] 期待値と分散

さて、正規分布のもつパラメータの推定を行いたいのだが、我々は時系列でない統計学においては、先に述べたi.i.d.に従っていると思われるいくつかのデータを抽出することで、**推定量から母数(パラメータ)を推定**する。

ここで考えなければならないのは、時系列データの場合は、ある時点を固定した時、その**観測値は一つ**しか得られないということである(**SFの世界であればパラレルワールドが存在するかもしれないが！**)。先の日経平均の終値であれば、**2000年4月1日の観測値は一つしか存在しない。**このことを**標本系列の唯一性**と呼ぼう。このことが時系列の分析を決定的に難しくしている要因の一つだと言って良い。

さて、考えにくことではあるが、もしある系列が時間順序に並んでいながらも、同一の確率分布(i.i.d.)から発生していたと考えたらどうだろう。これならば、我々の普段行なっている推定という行為をためらいなくできるであろう。のちに詳しく述べるが、この考え方を**定常性(stationarity)**という。

1.2.3 時間による変化

大変な困難を伴う時系列分析だが、我々が適切な確率モデルを設定し、それが現実と合致し、ある程度の期間は予測も上手くいったとする。しかしながら、さらなる困難が我々を待ち受けている。それは、もしある期間は上手くいったとしても、時間によって、その**確率モデルが徐々に(あるときは急に)変化してしまう可能性**があるからだ。一般に我々が時系列データとして観測しようとしている対象は、このような構造を持っていることが多い(だからこそ時系列で記録しようとする)。**時系列モデルの寿命は短い。**このような問題に対し、様々な解決が試みられているが現実には観測にノイズが混入することから、完全にそれ(時系列モデルの寿命)を把

握するのは難しい問題である。一つの有力な方法に異常検知技術を応用した、**変化点検知(change detection)**の技術を挙げることができよう。cf.http://bookclub.kodansha.co.jp/product?isbn=9784061529083 (http://bookclub.kodansha.co.jp/product?isbn=9784061529083)

このように、時系列の統計解析は、一般のデータ解析とは様々な点で異なっている。このような特殊性が時系列解析を難しくさせているのである。