时间序列

1.常用统计学参数

一组数据需要观察的话，我们需要了解一下他们的组成结构，正如我们要了解原子、分子、电子等的结构一个道理。

以 Z_t 表示一组数据，或**一个时间序列。**

（一）通用的几个基本概念：均值、方差、标准差、协方差、相关系数

- 均值

均值（期望）是统计学中最常用的统计量，用来表明数据集中相对集中较多的中心位置。

数学表示： $u_t = E(Z_t)$

- 方差

方差是用来度量一组数据的离散程度。概率论中方差用来度量随机变量和其期望（即均值）之间的偏离程度。统计中的方差（样本方差）是每个样本值与全体样本值的平均数之差的平方值的平均数。

数学表示： $\delta_t^2 = E(Z_t - u_t)^2$

- 协方差

协方差用来度量两个变量各个维度偏离其均值的程度，这与只表示一个变量误差的方差不同。协方差的值如果为正值，则说明两者是正相关的(从协方差可以引出“相关系数”的定义)，结果为负值就说明负相关的，如果为0，也是就是统计上说的“相互独立”。

数学表示: $\operatorname{cov}(Z_{t1}, Z_{t2}) = E(Z_{t1} - u_{t1})(Z_{t2} - u_{t2})$

假设有两个随机变量X、Y，大致上有：

(1)若协方差为正数：X增大，Y增大；X减小，Y减小，即变化趋势相同。

(2)若协方差为负数：X增大，Y减小；X减小，Y增大，即变化趋势相反。

(3)若协方差为零：X与Y的变化没有任何关系。

- 相关系数

相关系数是研究变量之间线性相关程度的量。求出协方差之后，我们考虑一个问题就是协方差对应这每一个“协”关系，他们对应得比值是多少，所谓对应的比值可以理解为每一个“协”距离整体的距离比值是百分之几？两个的“协”对应他们的整体距离的比值是百分之几就能够表示他们之间有多相关，这个相关系数越大，表示这两个数值越有关系。可以理解为，如果两个序列，一个是3000多这个基数去变动，一个是10000多这个基数去变动，他们的绝对数据肯定是不一样的，但是他们的变动比率是一样的，所谓相关性也可以理解为把两个值统一化，在同一个维度来评价这两个值的协方差关系，因此在同一个维度来衡量这两个值的协方差关系就叫做相关系数。

数学表示： $r(Z_{t_1}, Z_{t_2}) = \dfrac{\operatorname{cov}(Z_{t_1}, Z_{t_2})}{\sqrt{\delta_{t1}^2}\sqrt{\delta_{t2}^2}}$

相关系数的绝对值越大，相关性越强：相关系数越接近于1或-1，相关度越强，相关系数越接近于0，相关度越弱。通常情况下通过以下取值范围判断变量的相关强度：

(1) 0.8–1.0 极强相关

(2) 0.6–0.8 强相关

(3) 0.4–0.6 中等程度相关

(4) 0.2–0.4 弱相关

(5) 0.0–0.2 极弱相关或无相关

时间序列的特点是一维，因此如果借用统计学上面的指标衡量，有些不太适宜。根据时间序列的特点，形成了自协方差、自相关函数、偏自相关函数。看到前面都加了一个“自”，原因是时间序列没法在找到一个别的数据和自己来进行比较；只能自己和自己来比较，自己和自己慢几拍（滞后期）的这些数据进行比较，所以加入了一个“自”.

（二）时间序列自有的几个基本概念：自协方差、自相关系数、偏自相关系数

1、自协方差

在统计学中，特定时间序列或者连续信号 Z_t 的自协方差是信号与其经过时间平移的信号之间的协方差。

数学表示： $r(k)=\frac{1}{n}\Sigma_{t=k+1}^{n}\left(Z_t-\bar{Z}\right)\left(Z_{t-k}-\bar{Z}\right)$

可以认为自协方差是某个信号与其自身经过一定时间平移之后的相似性，自协方差 $r(k)$ 就表示了在那个时延的相关性。

2、自相关系数（ACF）

自相关系数度量的是同一事件在两个不同时期之间的相关程度，形象的讲就是度量自己过去的行为对自己现在的影响。

数学表示： $ACF(k)=\Sigma_{t=k+1}^{n}\frac{\left(Z_t-\bar{Z}\right)\left(Z_{t-k}-\bar{Z}\right)}{\Sigma_{t=1}^{n}\left(Z_t-\bar{Z}\right)^2}$

自相关（autocorrelation），也叫**序列相关，是一个信号于其自身在不同时间点的相关度。非正式地来说，它就是两次观察之间的相似度对它们之间的时间差的函数。它是找出重复模式（如被噪声掩盖的周期信号）**，或识别隐含在信号谐波频率中消失的基频的数学工具。它常用于信号处理中，用来分析函数或一系列值，如时域信号。

3、偏自相关系数（PACF）

根据ACF求出滞后k自相关系数 $ACF(k)$ 时，实际上得到并不是Z(t)与Z(t-k)之间单纯的相关关系

因为Z(t)同时还会受到中间k-1个随机变量Z(t-1)、Z(t-2)、……、Z(t-k+1)的影响，而这k-1个随机变量又都和z(t-k)具有相关关系，所以自相关系数里面实际掺杂了其他变量对Z(t)与Z(t-k)的影响。

为了能单纯测度Z(t-k)对Z(t)的影响，引进偏自相关系数（PACF）的概念。对于平稳时间序列{Z(t)}，所谓滞后k偏自相关系数指在给定中间k-1个随机变量Z(t-1)、Z(t-2)、……、Z(t-k+1)的条件下，或者说，在剔除了中间k-1个随机变量Z(t-1)、Z(t-2)、……、Z(t-k+1)的干扰之后，Z(t-k)对Z(t)影响的相关程度。

数学表达：

$$PACF(k)=\frac{E\left(Z_t-EZ_t\right)\left(Z_{t-k}-EZ_{t-k}\right)}{\sqrt{E(Z_t-EZ_t)^2}\sqrt{E(Z_{t-k}-EZ_{t-k})^2}}=\frac{\operatorname{cov}\left[\left(Z_t-\bar{Z}_t\right),\left(Z_{t-k}-Z_{t-k}^{-}\right)\right]}{\sqrt{\operatorname{var}\left(Z_t-\bar{Z}_t\right)}\sqrt{\operatorname{var}\left(Z_{t-k}-Z_{t-k}^{-}\right)}}$$

总结：时间序列借用统计学的数据结构分析公式

（1）期望还是等与期望

（2）自协方差 = 协方差（期望用整个时间序列的期望，一个期望）

（3）自相关系数 = 相关系数（期望用整个时间序列的期望，一个期望）

（4）偏自相关系数 = 相关系数（期望用各自序列的期望，两个期望）

2.认识时间序列

什么是时间序列？

> 时间序列也称动态序列，是指将**某种现象的指标数值按照时间顺序排列而成的数值序列。**大量的社会经济统计指标都是依据年、季度、月、日，甚至实时（秒）统计的，因此，时间序列是某个统计指标（变量）长期变动的数值表现。

时间序列由两个组成要素构成：1、第一个要素是时间要素；2、第二个要素是**数值要素。**时间序列根据时间和数值性质的不同，可以分为时期时间序列和时点时间序列。例如下方某间食杂店的销售额时间序列，就是时期时间序列，统计的是每一年，在一年时间内该食杂店的销售总额。

年份	2012	2013	2014	2015	2016
销售收入(万元)	32	35	40	42.8	47

又如下方某家制造工厂的动能部门，表格数据表示锅炉的分时温度数据。每隔一个小时，系统自动记录一次锅炉的实时温度。可以发现，这里的温度数据是某个时间点的实时数据，所以该时间序列为时点时间序列。

时间	14:00	15:00	16:00	17:00	18:00
温度（C°）	185.4	187.3	186.6	186.9	185.2

时间序列可以反映某个现象的发展变化状态。通过对时间序列的分析，可以反映现象发展变化的趋势和规律，再通过对影响时间序列的各种因素进行测定，可以进一步解释现象变化的内在原因，为预测和决策提供可靠的数据支持。

时间序列分解

因为时间序列是某个指标数值长期变化的数值表现，所以时间序列数值变化背后必然蕴含着数值变换的**规律性**，这些规律性就是时间序列分析的切入点。一般情况下，时间序列的数值变化规律有以下四种：**长期变动趋势、季节变动规律、周期变动规律和不规则变动。**不同的数值变化规律是由不同影响因素决定的。这些影响因素有长期起作用的因素，也有短期因素；有可以预知和控制的因素，也有未知和不可控制的因素；这些因素相互作用和影响，从而使时间序列的变化趋势呈现不同的特点。根据影响因素对时间序列数值变化趋势的不同影响情况，可以分为四种影响因素：长期趋势影响因素、季节变动影响因素、循环变动影响因素和不规则变动影响因素。

长期趋势 T

长期趋势指的是统计指标在相当长的一段时间内，受到长期趋势影响因素的影响，表现出持续上升或持续下降的趋势，通常用字母 T 表示。例如，随着国家经济的发展，人均收入将逐渐提升；随着科学技术的发生，劳动生产率也不断提高。

季节变动 S

季节变动是指由于季节的转变使得指标数值发生周期性变动。由此可见，指标数值的季节变动是以年为周期的，一般以月、季、周为时间单位，不能以年作单位，通常用 S 表示。引起季节变动的因素有自然因素，也有人为因素。例如，蔬菜食品价格，棉衣销售量都会随着季节气温的变化而周期变化；每年的长假（五一、十一、春节）都会引起出行人数的大量增加。

循环变动

循环变动与季节变动的周期不同，循环变动通常以若干年为周期，在曲线图上表现为波浪式的周期变动。这种周期变动的特征变现为增加和减少交替出现。最典型的周期案例就是市场经济的商业周期。

不规则变动

不规则变动是由某些随机因素导致的数值变化，这些因素的作用是不可预知和没有规律性的，因此对数值的变化影响变形为不规则变动。

以上四种变动就是**时间序列数值变化的分解结果。**有时这些变动会同时出现在一个时间序列里面，有时也可能只出现一种或几种，这是由引起各种变动的影响因素决定的。正是由于变动组合的不确定性，时间序列的数值变化才那么千变万化。四种变动与指标数值最终变动的关系可能是**叠加关系**，也可能是**乘积关系。**

叠加模型：如果四种变动之间是相互独立的关系，那么叠加模型可以表示为：

$$Y = T + S + C + I$$

Y表示指标数值的最终变动；

T表示长期趋势变动；

S表示季节变动；

C表示循环变动；

I表示不规则变动；

乘积模型：如果四种变动之间存在相互影响关系，那么应该使用乘积模型：

$$Y = T * S * C * I$$

Y表示指标数值的最终变动；

T表示长期趋势变动；

S表示季节变动；

C表示循环变动；

I表示不规则变动；

反映在具体的时间序列图上，如果随着时间的推移，序列的季节波动变得越来越大，则反映各种变动之间的关系发生变化，建议使用乘积模型；反之，如果时间序列图的波动保持恒定，则可以直接使用叠加模型。

3.时间序列分析

时间序列分析分成两种形式：第一种是传统的时间序列分析方法，研究时间序列是否能被分解成上面介绍的四种变动，并解析引起每种变动的影响因素。第二种是时间序列的模型解析法，常用时间序列模型有**自回归（AR）模型、滑动平均（MA）模型、自回归滑动平均（ARMA）模型等。**

(1) 传统时间序列分析方法

① 长期趋势分析

②季节变动分析

③循环变动和不规则变动

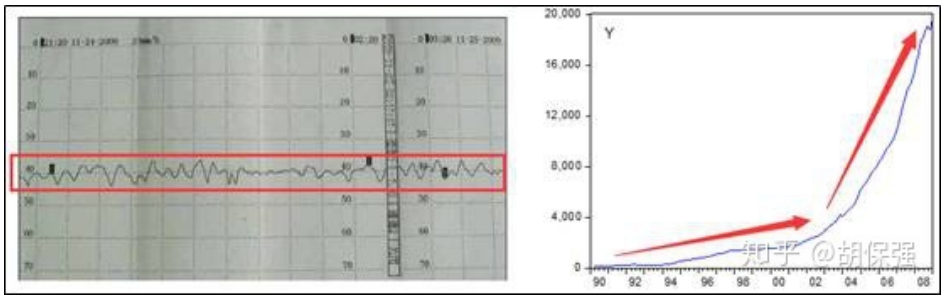
(2) 时间序列的模型解析法

后面的文章将重点介绍几种常用的时间序列模型：指数平滑模型、自回归（AR）模型、滑动平均（MA）模型、自回归滑动平均（ARMA）模型。文章将介绍这些模型的分析原理以及如何使用SPSS进行这些模型的时间序列分析。因为传统时间序列分析技术（时间序列分解法）的缺陷，所以统计学家开发出更为通用的时间序列分析方法，其中**AR/MA/ARMA/ARIMA**在这个发展过程中扮演了非常重要的角色，直到现在，它们都在实际工作生活中发挥重要作用。这四种分析方法的共同特点都是跳出变动成分的分析角度，从时间序列本身出发，力求得出前期数据与后期数据的量化关系，从而建立前期数据为自变量，后期数据为因变量的模型，达到预测的目的。来个通俗的比喻，大前天的你、前天的你、昨天的你造就了今天的你。

虽然AR/MA/ARMA/ARIMA是四种可以独立使用的分析方法，但是它们其实是互补的关系，适用于包含不同变动成分的时间序列。

①时间序列的平稳性

通俗介绍四种时间序列分析法之前，需要先回顾前面介绍的一个知识点，平稳时间序列(不存在趋势的序列，各观察值在固定水平上波动)和非平稳时间序列(包含趋势，季节性或周期性的序列，可包含一种或多种)，AR/MA/ARMA用于分析平稳时间序列，ARIMA通过差分可以用于处理非平稳时间序列。平稳时间序列和非平稳时间序列如下面两幅图所示：



左边的图是工业生产中的温度时间序列，它是围绕一个常数上下波动的，**也就是计算时间序列所有数值的平均值**，会等于这个常数。工业生产中对液面、压力、温度的控制过程；某地的气温变化过程；某条河流的水位变化过程基本都属于平稳时间序列。

右边的图是中国外汇储备额的时间序列，可以发现这个时间序列是有持续增长的，先慢后快，这是一个非平稳时间序列。在经济领域，例如一个国家的GDP、进出口额的时间序列基本都是非平稳时间序列。

一般具有长期趋势的时间序列都是非平稳时间序列。根据趋势的不同，可以使用差分将具有长期趋势的时间序列转换成平稳时间序列。例如，线性增长的长期趋势，可以通过一阶差分形成新的平稳的（消除长期趋势）时间序列：

例如，时间序列的数值为线性增长的(1,2,3,4,5,6,7,8)，经过一阶差分以后，新的时间序列的数值为(1,1,1,1,1,1,1)，就成为稳定的时间序列了。

根据长期趋势的发展趋势不同，可以进行差分的次数和方法也不相同，一般的规律如下：

- 一次差分的时间序列数值大体相同，配合直线趋势；
- 二次差分的时间序列数值大体相同，配合二次曲线
- 对数的一次差分的时间序列数值大体相同，配合指数曲线
- 一次差分的环比值大体相同，配合修正指数曲线
- 对数一次差分的环比值大体相同，配合Gompertz曲线
- 倒数一次差分的环比值大体相同，配合Logistic曲线

现在常用的时间序列模型分为两种，一种是基于统计学的线性模型，一种是神经网络模型。

②AR/MA/ARMA/ARIMA模型(线性模型)

这四种模型的名称都是它们英文全称的缩写。AR模型称为自回归模型（Auto Regressive model）；MA模型称为移动平均模型（Moving Average model）；ARMA称为自回归移动平均模型（Auto Regressive and Moving Average model）；ARIMA模型称为差分自回归移动平均模型。

- **AR模型**

如果某个时间序列的任意数值可以表示成下面的回归方程，那么该时间序列服从p阶的自回归过程，可以表示为AR(p)：

$x_t = \emptyset_1 x_{t-1} + \emptyset_2 x_{t-2} + \cdots + \emptyset_p x_{t-p} + u_t$
$x_t, x_{t-1}, x_{t-2}, \ldots\ldots, x_{t-p}$为不同时间点记录的指标数值；
$\emptyset_1, \emptyset_2, \emptyset_3, \ldots\ldots, \emptyset_p$为自回归系数；
u_t成为该时间序列的白噪声；
$x_t = \emptyset_1 x_{t-1} + u_t$称为 1 阶的自回归过程AR(1);
$x_t = \emptyset_1 x_{t-1} + \emptyset_2 x_{t-2} + u_t$称为 2 阶的自回归过程AR(2);

可以发现，**AR模型利用前期数值与后期数值的相关关系（自相关），建立包含前期数值和后期数值的回归方程，达到预测的目的，因此成为自回归过程。**这里需要解释白噪声，大家可以将白噪声理解成时间序列数值的随机波动，这些随机波动的总和会等于0，例如前面统计基础文章中介绍的，某条饼干的自动化生产线，要求每包饼干为500克，但是生产出来的饼干产品由于随机因素的影响，不可能精确的等于500克，而是会在500克上下波动，这些波动的总和将会等于互相抵消等于0。

- **MA模型**

如果某个时间序列的任意数值可以表示成下面的回归方程，那么该时间序列服从q阶的移动平均过程，可以表示为MA(q)：

$x_t = u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \emptyset_q u_{t-q}$
$u_t, u_{t-1}, u_{t-2}, \ldots\ldots, u_{t-q}$表示不同时间点的白噪声项；
$\theta_1, \theta_2, \theta_3, \ldots\ldots, \theta_q$为移动回归方程系数；
x_t表示时间点 t 对应的指标数值；

可以发现，某个时间点的指标数值等于白噪声序列的加权和，**如果回归方程中，白噪声只有两项，那么该移动平均过程为2阶移动平均过程MA(2)。**比较自回归过程和移动平均过程可知，移动平均过程其实可以作为自回归过程的补充，解决自回归方差中白噪声的求解问题，两者的组合就成为自回归移动平均过程，称为ARMA模型。

- **ARMA模型**

自回归移动平均模型由两部分组成：自回归部分和移动平均部分，因此包含两个阶数，可以表示为ARMA(p,q)，p是自回归阶数，q为移动平均阶数，回归方程表示为：

$$x_t = \emptyset_1 x_{t-1} + \emptyset_2 x_{t-2} + \cdots + \emptyset_p x_{t-p} + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \emptyset_q u_{t-q}$$

从回归方程可知，自回归移动平均模型综合了AR和MA两个模型的优势，在ARMA模型中，自回归过程负责量化当前数据与前期数据之间的关系，移动平均过程负责解决随机变动项的求解问题，因此，该模型更为有效和常用。

平稳性对于我们分析时间序列至关重要。如果一个时间序列不是平稳的，通常需要通过差分的方式将其转化为平稳时间序列。

对于一个时间序列，如何确定它是否满足平稳性要求？通常采用 ADF 检验。

- **ARIMA模型**

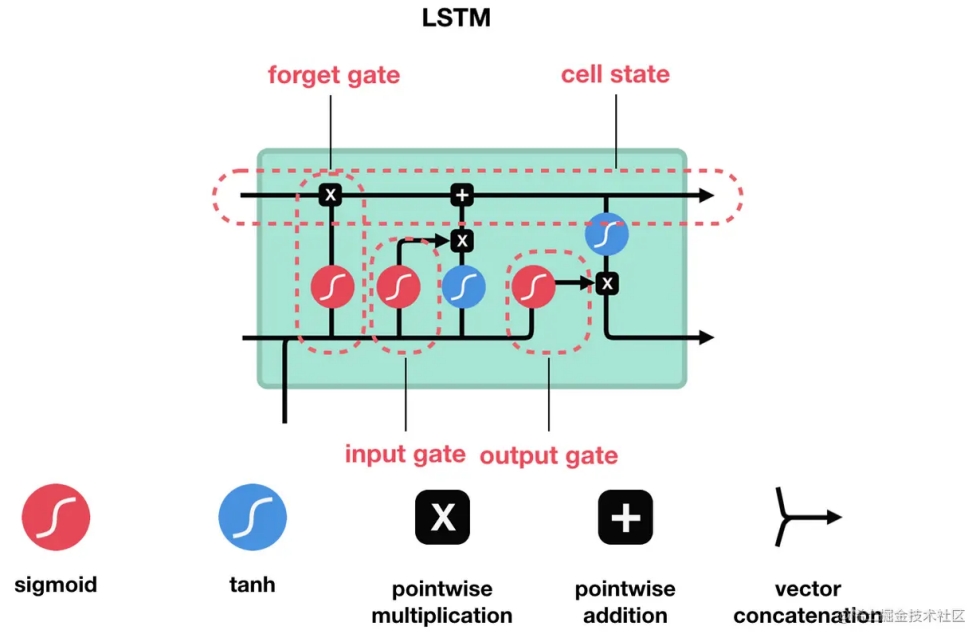
介绍时间序列平稳性时提到过，AR/MA/ARMA模型适用于平稳时间序列的分析，当时间序列存在上升或下降趋势时，这些模型的分析效果就大打折扣了，这时差分自回归移动平均模型也就应运而生。ARIMA模型能够用于齐次非平稳时间序列的分析，这里的齐次指的是原本不平稳的时间序列经过d次差分后成为平稳时间序列。

在现实生活中，存在很多非平稳的时间序列，它们的均值和方差是随着时间的变化而变化的，幸运的是，统计学家们发现，很多时间序列本身虽然不平稳，但是经过差分（相邻时间点的指标数值相减）之后，形成的新时间序列就变成平稳时间序列了。因此，差分自回归移动平均模型写成ARIMA(p,d,q)。p代表自回归阶数；d代表差分次数；q代表移动平均阶数。在spss软件中，有时输出的ARIMA模型包括6个参数：ARIMA(p,d,q)(P,D,Q)，这是因为如果时间序列中包含季节变动成分的话，需要首先将季节变动分解出来，然后再分别分析移除季节变动后的时间序列和季节变动本身。这里小写的p,d,q描述的是移除季节变动成分后的时间序列；大写的P,D,Q描述的是季节变动成分。两个部分是相乘的关系。因此，ARIMA(p,d,q)(P,D,Q)也被称为复合季节模型。

③ 非线性模型LSTM

这里主要总结一下长短期记忆网络（LSTM）的原理，LSTM是根据循环神经网络（RNN）改进而来，主要是为了解决RNN中出现的梯度消失问题以及序列过长时会丢失较早期的数据的情况。其主要的内部机制是通过三个门调节信息流，了解序列中哪些数据需要保存或者丢弃。

核心概念：LSTM 的核心概念是细胞状态，三个门和两个激活函数。细胞状态充当高速公路，在序列链中传递相关信息。门是不同的神经网络，决定在细胞状态上允许那些信息。有些门可以了解在训练期间保持或忘记那些信息。



上图为一个LSTM的cell单元，其中他的输入段的上层是 c_{t-1} 为上一个单元的状态，下层为 h_{t-1} 即上一个单元输出的隐藏状态。而经过整个单元之后上层输出这个单元的状态 c_t ，下层输出这个单元的隐藏状态 h_t 。