Statistical and Application 统计学与应用, **2014, 3, 193-202**
Published Online December 2014 in Hans. http://www.hanspub.org/journal/sa
http://dx.doi.org/10.12677/sa.2014.34026

The Comprehensive Utilization of Correlation Coefficient Information in the Empirical Study

Jianjun Wang, Huiping Yang

School of Statistics and Information, Xinjiang University of Finance and Economics, Urumqi
Email: XJWJJ@XJUFE.EDU.CN

Received: Oct. 20th, 2014; revised: Nov. 22nd, 2014; accepted: Dec. 2nd, 2014

Abstract

Correlation coefficient is a very important concept in statistics. Most purposes of statistics are to research and find relevant information between variables. Correlation coefficient not only provides the related degree and direction. More importantly, it can get useful auxiliary information by correlation. This paper describes how to make use of correlation coefficient information in empirical research. It analyzes occasion and method which use auxiliary information provided by correlation coefficient from multiple perspectives. For an overview of the role of correlation coefficient, how to avoid spurious correlation is of great significance.

Keywords

Correlation Coefficient, Auxiliary Information, Spurious Correlation

相关系数信息在实证研究中的综合利用

王建军，杨辉平

新疆财经大学统计与信息学院，乌鲁木齐
Email: XJWJJ@XJUFE.EDU.CN

收稿日期：2014年10月20日；修回日期：2014年11月22日；录用日期：2014年12月2日

摘　要

相关系数是统计学中非常重要的概念，统计学经常是要研究相关，寻找相关，相关系数不仅提供了相关程度与方向，更重要是利用相关关系得到辅助信息。本文阐述如何在实证研究中综合利用相关系数信息，多角度分析了利用相关系数所提供的辅助信息的场合与方法，对全面认识相关系数，如何避免伪相关具有重要意义。

关键词

相关系数，辅助信息，伪相关

1. 引言

相关系数是统计学中一个非常重要的概念，是用于测定两个变量之间线性相关程度和相关方向的统计分析指标。维基百科指出，相关系数的意义是用来衡量两个变量相对于其相互独立关系的距离。

目前统计学教材[1]之中只介绍相关系数计算公式及性质，缺少如何应用相关关系解决实际问题，相关系数有多少种类？每一类相关系数适合分析哪一类问题？不同的相关系数结果怎么解释？使用相关系数进行实证研究分析时需要注意什么？当从理论上分析变量之间应该存在相关关系，而样本数据不相关或相关性不强时，又如何找出真实的相关性关系？本文从相关系数的信息综合利用角度讨论相关系数在实证研究中的各种利用的方法。

2. 相关关系与相关系数

2.1. 相关关系与相关系数的联系与区别

相关关系与相关系数是两个有联系但也有区别的重要统计概念。

相关关系是一种陈述，是指两个变量从实践经验和理论分析上确实存在某种联系，但是这种关系并不是确定的一一对应关系。如从实践经验可得出，一个人的身高 X 越高，衣袖 Y 一般越长，但同样身高的人穿衣尺寸并不同，也存在身高低而衣袖长的人；从经济理论分析，居民收入越高，储蓄额会越大，但确实存在收入下降但储蓄额却上升的情况。这两个实例表示的关系就是相关关系。这种两个变量之间在数量上非确定性的对应关系称为“相关关系”。人们在实践中会经常发现了一个变量的变化会引起另一个变量的变化，这种相关关系的发现可以帮助找到某些统计规律。为了验证、研究这些规律，需要有意识的进一步通过大量试验、观察，搜集两个变量之间的对应数值，对这两个变量的一系列观测值进行统计分析。如回归分析方法，形成具有一定概率的统计规律。函数关系是相关关系的特例。相关关系可以分为线性相关和非线性相关。

相关系数一般指皮尔逊相关系数，是一个用于描述和衡量变量之间线性相关程度与方向的统计量。具体来说相关关系和相关系数的区别和联系如表 1：

由表 1 可知，相关系数只度量了变量间相关关系中的一部分，而不是全部。

相关系数的分类。根据研究对象的变量个数不同，相关系数可以分为分析一个变量与一个变量相关的简单相关系数，分析一个变量与一群变量相关的复相关系数；分析一群变量与一群变量相关的典型相关系数(CCA)。根据变量特性分析两个分类变量(定类或定序)的相关关系的列联系数(contingency coefficient)；利用变量的秩(rank)和协同(concordant)计算的 Spearman 和 Kendall 等非参数相关系数等[2] [3]。

Table 1. The difference and connection between correlation and correlation coefficient
表 1. 相关关系与相关系数的区别与联系

相关关系	相关系数(Pearson)
从经验和理论上分析出有关系，用文字表述两变量间有一定的关系。	定量衡量两变量间线性相关程度与方向
可以是线性和非线性相关关系	只能用于描述线性关系
是计算相关系数的前提	是分析线性相关关系的工具
理论上无相关关系的变量是独立的	相关系数为 0 不一定独立
相关关系的散点图要呈现一定的规律性	散点图呈椭圆状
有关系的变量可以是数值型与分类型变量	主要用于连续型数值变量

2.2. 独立性与相关性的关联

根据概率论和数理统计中的定义，如果两个随机事件是独立的那么这一事件的发生不会影响到另一事件发生的概率。设 A、B 表示两个事件，如果满足：

$$P(A|B)=P(A),\ P(B|A)=P(B),\ P(AB)=P(A)P(B)$$

我们则称 A、B 事件是相互独立的。类似的，如果 A、B 两个事件满足：

$$P(A|B)\neq P(A),\ P(B|A)\neq P(B)$$

上式表明 B 事件的发生对 A 事件发生的概率是有影响的，同样，B 事件的发生对 A 事件发生的概率也是有影响的，那么我们称 A、B 事件是有关联的或相关的。

统计学中两个变量独立就不相关，不独立就有相关关系。在数学理论研究中经常假设变量之间相互独立的条件，此时的联合分布密度等于边缘分布密度的乘积，由此可以很容易得到一些看似很精彩的数学结论。实际上在现实生活中完全独立变量很少需要放在一起研究，这时变量之间相关性的分析就显得尤为重要。统计学家主要研究相关变量，可以得到有用的信息。

3. 相关系数的定义与性质

3.1. 相关系数定义

3.1.1. 参数统计的相关系数

相关系数又叫皮尔逊相关系数或线性相关系数，这是根据样本数据计算的度量两个变量之间线性关系强度的统计量[1]，被定义为协方差除以其标准差的乘积。其计算公式为：

$$\text{总体相关系数}\ \gamma_{XY}=\frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}=\frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

其中，$\operatorname{cov}(x,y)=\sigma_{xy}$ 为协方差，σ_X 和 σ_Y 是标准差。

$$\text{样本相关系数}\ r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{s_{xy}}{s_x s_y}$$

其中 n 为样本量，若将两个变量的样本值看成 n 维向量，当均值 $\bar{x}=\bar{y}=0$ 时，得到与夹角余弦相同的公式。

$$r_{xy}=\frac{\sum_{i=1}^{n}(x_i y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2\sum_{i=1}^{n}(y_i)^2}}$$

由于两个随机变量组成随机向量的协方差矩阵是正定的对称矩阵

$$|\Sigma|=\begin{vmatrix}\sigma_{xx} & \sigma_{xy}\\ \sigma_{yx} & \sigma_{yy}\end{vmatrix}=\sigma_{xx}\sigma_{yy}-\sigma_{yx}\sigma_{xy}=\sigma_{xx}\sigma_{yy}\left(1-\frac{\sigma_{yx}^2}{\sigma_{xx}\sigma_{yy}}\right)\geq 0$$

即 $\sigma_{xx}\sigma_{yy}\left(1-r^2\right)\geq 0$

所以 r 的取值范围在–1 到 1 之间，若 $r>0$，表示两个变量存在正相关关系，即一个变量的值与另外一个变量值同方向变化；若 $r<0$，表示两个变量存在负相关关系，即一个变量的值与另外一个变量值反方向变化；其缺点是容易受极端值的影响，所以需要引入非参数的具有稳健性相关系数。

在统计学中，一般采用以下 t 统计量对相关系数的相关程度进行检验：

$$t=(n-2)^{1/2}\left(\frac{r^2}{1-r^2}\right)^{1/2}$$

3.1.2. 非参数统计的相关系数

Pearson 相关系数不稳健且要求双变量服从于正态分布的连续型变量，然而现实中大部分变量却并不服从于正态分布，这时候采用简单相关系数来度量相关关系并不合适，应当采用非参数 Spearman 和 Kendall 相关系数来进行度量，具体计算及应用条件如下：

Spearman 相关系数[2] [3]，也称秩相关(rank correlation)，设有两个变量 X 与 Y，R 是 X 的秩，Q 是 Y 的秩，则相关系数为

$$r_s(R,Q)=\frac{\sum\left(R_i-\frac{1}{n}\sum R_i\right)\left(Q_i-\frac{1}{n}\sum Q_i\right)}{\sqrt{\sum\left(R_i-\frac{1}{n}\sum R_i\right)^2\sum\left(Q_i-\frac{1}{n}\sum Q_i\right)^2}}$$

Spearman 相关系数适用于度量连续性、离散型包括普通变量的相关程度，且不易受极端值影响，具有良好的稳健性，建议最好用 Spearman 相关系数进行实证研究。在统计学中，一般采用以下 t 统计量对 Spearman 相关系数的相关程度进行检验：

$$t=\frac{r_s-\mu_{r_s}}{s_{r_s}}=\frac{r_s-0}{\sqrt{\frac{1-r_s^2}{n-2}}}=r_s\sqrt{\frac{n-2}{1-r_s^2}}\sim t(n-2)$$

Kendall 相关系数适用于度量不服从双变量正态分布或总体分布未知的等级资料，变量 X 与 Y 的 Kendall 相关系数定义为：

$$r=\frac{C-D}{\frac{1}{2}n(n-1)}=\frac{4C}{n(n-1)}-1$$

其中 n 为样本量，以 C 为协同(concordant)对数，D 为不协同对数，即将所有的项目按照第一个变量 X 排序后，第二个变量 Y 的排秩能保持与变量 X 一致的大小顺序的数对的个数之和，分母 $n(n-1)/2$ 可以解释

为总对数。Kendall τ 的主要优点是它的分布能非常快的接近于正态分布，且当 X 和 Y 独立的零假设为真时，Kendall τ 的正态逼近比 Spearman 的要好[3]。

3.2. 相关系数的性质

1) 若存在常数 a，b，使得 $Y=a+bX$，则 $|r_{XY}|=1$。表明当相关系数等于 1 时，变量 X 与 Y 存在完全线性关系。当相关系数等于 0 时，则称变量 X 与 Y 不相关。

若 $Y=a+bX$，假设 $E(X)=\mu$，$D(X)=\sigma^2$ 则 $E(Y)=a+b\mu$，$D(Y)=b^2\sigma^2$

$$E(XY)=E\left(aX+bX^2\right)=a\mu+b\left(\mu^2+\sigma^2\right)$$

$$Cov(XY)=E(XY)-E(X)E(Y)=b\sigma^2$$

所以当 $b\neq 0$ 时，则 $r_{XY}\neq 0$，当 $b=0$ 时，则 $r_{XY}=0$

2) 决定系数 R^2 具有线性不变性。若 A，B 可逆，则有 $r(Ax,By)=r(x,y)$。相关系数相同，方差和协方差不同。

3) 相关系数中变量的地位同等。相关系数不考虑因变量与自变量。

4) 标准化变量的协方差就是相关系数。标准化后，标准差为 1，相关系数等于协方差除以标准差所以仍等于协方差。标准化后均值为 0，相关系数就是夹角余弦。

5) 相关系数易受极端值影响。由于用到均值，而平均数易受极端值影响，相关系数也易受极端值影响。

6) 比较两个相关系数必须样本量相等。不能随意比较两个相关系数，样本量相等才可比较。相关系数的检验结果与样本量有关系。

4. 相关系数信息在实证研究中的应用

4.1. 相关系数信息在统计分支中应用

在现实社会经济生活中，各种社会经济因素相互作用，关系极其复杂，为了快速找到哪些因素之间存在着相互影响，此时相关系数就显得尤为重要。表 2 列出相关系数被广泛应用于统计学各个分支：

4.2. 相关系数在定性变量的关联分析中应用——列联系数

设两个定性变量(分类或有序变量)X 与 Y，X 分为 K 类，Y 分为 m 类，分类 X 与 Y 的关联关系，需要

Table 2. Correlation coefficient used in statistical branch
表 2. 相关系数在统计学中分支的应用

统计分支	相关系数应用
概率论	定义相关系数概念与计算公式
描述统计	相关系数计算、性质，相关程度与方向
多元统计分析	变量聚类分析，主成分和因子分析压缩维度，典型相关
回归分析	利用变量相关建立模型，模型拟合度，偏相关系数来筛选变量
计量经济学	检验模型自相关，共线性，因果关系
时间序列	利用自相关建立模型
非参数统计	Spearman，kendall 相关系数，列联表变量相关
抽样调查	利用相关变量作为辅助信息减少估计误差

检验是否独立。列联表检验时常用拒绝独立就存在相关性。

$$\chi^2=\sum_{j=1}^{k}\sum_{i=1}^{m}\frac{\left(f_{ij}-np_{i.}p_{.j}\right)^2}{np_{i.}p_{.j}}\sim\chi^2\left[(m-1)(k-1)\right]$$

例如为了研究一个人在计算机行业从事的职业类型于他在该行业中的从业年限是否存在关系，调查了 250 人。调查结果如表 3：

首先建立零假设：一个人在计算机行业从事的职业类型与他在该行业中的从业年限相互独立。利用 SPSS 计算可得 Pearson 卡方值为 65.66，对应的 P 值为 0.000 小于显著性水平 0.05，这属于小概率事件，所以可以认为一个人在计算机行业从事的职业类型与他在该行业中的从业年限不相互独立，存在相关关系。

上面所求的 χ^2 统计量的值与列联表的行列数目和样本量有关，为了消除对样本大小的依赖，我们可以使用 Pearson 的列联系数(相关系数)：

$$\varphi=\sqrt{\frac{\chi^2}{\chi^2+n}}$$

其中 χ^2 为卡方统计量值，n 为样本量。

4.3. 相关系数在回归模型分析中应用

1) 有益于模型的相关—因变量与自变量的相关

回归模型主要分析因果关系，一般所说的回归模型指的是线性回归模型，只有存在线性相关的变量才能建立回归模型，相关是回归模型的基础。散点图能够从直观上观察两个变量之间是否存在线性相关关系，而相关系数能够定量的测度两个变量线性相关关系程度。变量之间高度的相关关系有助于提高模型的拟合优度，减少估计的误差。

2) 有损于模型的相关。有些变量的相关在回归模型中是有损于模型拟合的。如：

①自变量的共线性(自变量相关)

回归模型假设自变量间不相关，多元回归分析中变量的高度相关可引起共线性[4]，有文献指出相关系数 0.8 以上表示存在严重的共线性。

$$\beta=\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}Y$$

上式只有在逆矩阵存在的情况下才可以估计出参数 β，X 的相关性可能导致上述矩阵的逆不存在，影响模型参数估计。这时候可以通过岭回归和主成分回归的方法构造一个可逆矩阵解决。

$$\beta=\left(X^{\mathrm{T}}X+K\right)^{-1}X^{\mathrm{T}}Y$$

Table 3. Computer industry career types and working years
表 3. 计算机行业的职业类型与从业年限

		职业类型			
		经理	程序员	操作员	系统分析员
从业年限	0~2 年	6	41	11	13
	3~5 年	28	16	23	24
	5 年以上	47	10	12	19

②自变量与误差项相关

回归模型的一个基本假设是每一个自变量与误差项相互独立，如果这个假设不成立，普通最小二乘法回归参数估计量将不再是无偏和一致的。这时候可以通过工具变量法加以解决，具体来说将测量误差用工具变量替代，工具变量是一个与自变量 X 高度相关，同时与方程的误差项不相关的新变量 Z。

③误差项自相关

回归模型中残差是假设不相关的，经过 DW 检验若残差之间存在显著自相关，即 $\text{cov}\left(e_i,e_j\right)\neq 0$，则说明还有重要变量被包含于残差中，所建的模型仍有重要信息没有被表述出。误差项之间自相关系数的计算如下：

$$\rho=\frac{\text{cov}\left(\varepsilon_t,\varepsilon_{t-1}\right)}{\sigma_t^2}=\frac{\text{cov}\left(\varepsilon_t,\varepsilon_{t-1}\right)}{\left[\text{var}\left(\varepsilon_t\right)\right]^{1/2}\left[\text{var}\left(\varepsilon_{t-1}\right)\right]^{1/2}}$$

3) 回归系数与相关系数的关系

当把数据进行标准化后，回归系数与相关系数相等。例如一元回归模型 $y=\beta_0+\beta_1x$，由于标准化数据中 $\sigma_X=1$，$\sigma_Y=1$，则

$$\gamma_{XY}=\frac{\text{cov}\left(X,Y\right)}{\sqrt{\text{var}\left(X\right)\text{var}\left(Y\right)}}=\frac{\beta_1}{\sigma_X\sigma_Y}=\beta_1$$

4) 决定系数与相关系数的关系

一元回归模型 $y=\beta_0+\beta_1x$ 中，决定系数的计算如下：

$$r^2=\frac{\left(\dfrac{\sigma_{xy}^2}{\sigma_x^2}\right)}{\sigma_y^2}=\frac{\beta_1^2\sigma_x^2}{\sigma_y^2}$$

由上式，决定系数的含义可以解释为用标准差做计量单位，X 变动一个标准差时，Y 变动 r 个标准差。或者说由 X 解释 Y 变动的部分，在 Y 方差中所占的比例。由相关系数就可以求出 Y 对 X 回归模型的决定系数。

4.4. 利用相关系数提高抽样估计精度

在抽样调查中，并不是只调查一个变量，而是需要同时调查很多与样本有关的变量，而这些变量之间往往存在较强的相关关系，这时候充分利用相关性这一非常重要的信息，选用与调查变量有较密切的正相关关系的变量作为辅助变量，可以减少一些抽样估计误差，提高估计精度。

$$V\left(\bar{y}_{lr}\right)\approx\frac{1-f}{n}S_y^2\left(1-\rho^2\right)$$

其中ρ为y与x的相关系数，由上式可以看出抽样估计误差与相关系数有关；简单估计量和比率估计量都可以看成是 $\beta=0,\beta=\dfrac{\bar{y}}{\bar{x}}$ 回归估计量的特例。

4.5. 利用相关系数对变量分组

1) 相关系数与因子分析

因子分析是一种降维技术和变量分类方法，其目的是用有限个不可观测的潜变量来解释原始变量之间的相关关系。因子分析本质上是建立在变量相关的基础之上，因子分析利用相关性将变量分组，原始变量间的相关系数可用于说明因子分析的必要性，一些相关系数高的变量是由同一个因子这个潜在的不

可观测的变量决定的。

$$\underset{p\times 1}{X}=\underset{P\times m}{A}\underset{m\times 1}{f}+\underset{p\times 1}{e}$$

其中 F 为因子向量，A 为载荷矩阵。一般需要进行因子旋转才能得到明确的变量分组。原始变量与因子的关系由相关系数表示，因子载荷就是 X 与 F 的相关系数，相关系数值越大说明因子对这个原始变量的解释程度越高。

2) 相关系数与变量的聚类分析

聚类分析(cluster analysis)，变量聚类是将相近的、相关程度高的变量合并为同一类达到对变量分组的目的。聚类中相关系数是一种测试相似程度的方法，样本聚类一般用距离，越小则越近，变量聚类用相关系数，正好与距离相反，相关系数越大表示越近，将相关程度高的合并为一类。

4.6. 利用相关系数进行维度压缩

1) 主成分分析。将变量的相关视为信息冗余、重复，此时的相关是负面的，若 P 维空间的变量高度相关，说明不需要这么多的维度，可以用较少的维数的主成分分析的现象，这就是降维技术，主成分利用线性变换，将相关变量变换为不相关的变量—主成分，进行数据压缩-降维。

$$Z=a_1x_1+a_2x_2+\cdots+a_px_p=Ax$$

主成分分析的本质是将所有变量投影到方差最大的方向，提取最主要的信息。

2) 典型相关分析。典型相关(Canonical Correlation Analysis)是研究两组变量之间相关性的一种统计分析方法。为了从总体上把握两组指标之间的相关关系，分别在两组变量中提取有代表性的两个综合变量 U_1 和 V_1(分别为两个变量组中各变量的线性组合)，利用这两个综合变量之间的线性相关关系来反映两组指标之间的整体相关性。

$$u_1=l_1'x,\quad v_1=m_1'y$$

其中 $x=\left(x_1,x_2,\cdots,x_p\right)'$，$y=\left(y_1,y_2,\cdots,y_p\right)'$，$l_1'$、$m_1'$ 为常数列向量，找到使 u_1 和 v_1 的相关系数 $\rho\left(u_1,v_1\right)$ 达到最大的 l_1' 和 m_1'。

4.7. 利用自相关系数进行时间序列预测分析

时间序列自相关函数(autocorrelation function)，简记为 ACF

$$\rho(t,s)=\frac{E\left(X_t-\mu_t\right)\left(X_s-\mu_s\right)}{\sqrt{DX_t\cdot DX_s}}$$

自相关函数用于度量同一事件在两个不同时期之间的相互影响的程度，时间序列建立在序列自相关基础之上，比如 ARIMA 模型。

4.8. 利用相关系数检验调查问卷的信度

信度系数是相关系数的一类。信度分析指在社会测量中，采用同样的方法对同一对象重复进行测量时，其获得结果的一致性程度[5]。为了定量研究某些特定的问题而数据又难以获取时，我们经常采用问卷调查的方式进行，然而问卷的调查结果是否可信？问卷设计的是否足够合理？这时可以采用 Cornbach's Alpha 检验对样本数据进行可信度检验，$\alpha=\frac{k}{k-1}\left|1-\frac{\sum_{i=1}^{k}\sigma_i^2}{\sum_{i=1}^{k}\sigma_i^2+2\sum_{i=1}^{k}\sum_{j<1}^{k}\sigma_{ij}}\right|$，其中 k 为所探讨问卷项目个数。

样本的内部一致性检验 Cronbach alpha 系数不少于 0.7，说明样本通过了内部一致性检验。信度的大小可以用这两次测量结果的相关系数来表示。

5. 相关系数使用误区

1) 相关系数为 0 不一定独立。可能存在非线性相关关系。

相关与独立，独立一定不相关，但不相关(相关系数为 0)，协方差为 0，不一定独立。我们知道只有线性相关才能利用相关系数和回归模型处理这些信息。如何将不独立，但相关系数为 0(不相关)的两个变量使其有线性相关呢？使其化为直线相关，以便利用回归模型。

首先从理论上分析是否独立，或从经验分析是否有关系。其次对理论上有相关关系的变量，但相关系数很弱或不理想，就需要用曲线形式描述，需要做变换。非线性回归的核心是将非直线关系，通过变换成为直线关系。

取对数影响相关程度吗？一般变量取对数后能增加相关程度，

2) 相关关系不一定是因果关系。因果关系一定相关(但不一定是直线相关)。相关关系仅仅表示两个变量之间存在某种关系，究竟是谁影响谁并不确定。因果关系可以确切的知道一个变量影响另一个变量，因果关系的确定需要经济理论的支撑，做预测是不需要有因果关系的。

3) 一般不能仅根据相关系数大小比较相关程度。比较两组相关系数时，只有当样本容量相等时才可以比较相关系数大小，相关系数小就相关程度低；当样本容量不相等时，相关系数(绝对值)大不一定相关程度高。当样本容量不相等时，这时相关系数的大小和相关程度的检验结果可能不一致，可能出现相关系数大而相关检验的显著性水平低。统计检验表明，相关系数显著性与样本容量有关。

例：为了探究青少年的历史与语文成绩、身高与体重是否存在一定的关联，现随机抽取某中学高三年级 1 班学生各 6 名，历史：66, 68, 69, 70, 75, 78；语文：67, 79, 80, 82, 82, 95；

2 班 20 名身高：150, 158, 160, 163, 166, 167, 169, 170, 172, 175, 175, 175, 180, 182, 182, 185, 185, 186, 186, 189；体重：50, 50, 72, 55, 65, 89, 67, 67, 70, 90, 75, 75, 80, 80, 92, 82, 82, 96, 96, 100。

经过计算，变量 X_1 与 X_2 的相关系数为 0.879，对应的双侧检验的 $P = 0.021$，在显著性水平 0.01 下不显著；Y_1 与 Y_2 的相关系数为 0.844，对应的双侧检验的 $P = 0.000$，这时相关系数(绝对值)大的对应的双侧检验的显著性水平反而低，那么谁的相关程度高呢？实际上，检验统计量 T 的结果与样本容量有关，样本量不相等时这样的比较不可取！

4) 伪相关现象(Spurious correlation)

在实际应用中，当我们计算两个完全没有任何关系的变量相关系数时，有时候得到的相关系数较大，而且经过检验是显著不为 0 的，统计上将这种现象称之为伪相关。伪相关现象是由于变量之间都存在某种相同的变化趋势，或者说存在着第三个变量将他们联系在一起，潜在变量的存在(潜在变量的影响)，两个变量 X，Y 都受某个潜在变量 Z 的影响导致共同反应(common response)。

两个经济变量之间的高度相关关系，有时并不是这两个经济变量本身的内在联系所决定的，它完全可能由另外一个变量的媒介作用而形成高度相关；忽略了媒作用，理论上为负相关的变量可能得到正相关关系，时间序列不平稳时常常出现伪相关。伪相关的存在经常让我们得到一些看似合理实际上错误的信息，这时候必须找出并消除潜在变量的影响，两变量之间的真正关系才能浮出水面。所以，我们绝不能只根据相关系数很大，就认为两者经济变量之间有直接内在的线性联系。此时要准确地反映两个经济变量之间的内在联系，就不能简单的计算相关系数，而是需要考虑偏相关系数，偏相关关系则是在扣除或固定某两个变量以外的其他变量对它们的影响以后,这两个变量之间的相关关系,它反映了事物间的本质联系[6]。

6. 结论与启示

从以上分析可以看出，相关关系与相关系数是两个有联系但也有区别的重要统计概念。相关就是用于研究和解释两个变量之间存在相互关系的，相关关系可以是非线性相关或直线相关；相关系数是一个用于描述和衡量变量之间线性相关程度的统计量。对于存在非线性关系的变量可以通过恰当的变化转换为线性关系。

相关系数主要分为简单相关系数，非参数相关系数，简单相关系数要求数据服从于正态分布，当数据不服从于正态分布时，应适用非参数相关系数进行度量。相关系数广泛应用于列联表、回归模型、抽样调查、多元统计、时间序列等学科中，然而在应用中必须要分清楚有益的和无益的相关，有益的相关可以在分析问题之前得到一些非常有用的辅助信息，而无益的伪相关却会误导得到一些错误的结论，在使用相关系数分析问题时必须注意它的使用条件，避免使用陷阱。

参考文献 (References)

[1] 贾俊平，何晓群，著 (2012) 统计学. 第五版，中国人民大学出版社，北京, 269-270.

[2] 王星编，著 (2009) 非参数统计. 清华大学出版社，北京, 181-182.

[3] 崔恒建，译 (2006) 实用非参数统计. 人民邮电大学出版社，北京, 230-231.

[4] 宋廷山 (2008) 相关系数统计量的功能及其应用探讨. *统计教育*, **11**, 27-31.

[5] 黄润龙 (2010) 数据统计分析-SPSS 原理及应用. 高等教育出版社，北京, 184-185.

[6] 王海燕，杨方廷 (2006) 标准化系数与偏相关系数的比较与应用. *数量经济技术经济研究*, **6**, 150-155.