

frilly / December 10, 2024 10:07AM

[&quot;相關性分析&quot;在資料分析中的應用](#)

- [1. 1.01 什麼是相關性分析？](#)
- [2. 2.02 相關性分析實際案例](#)
- [3. 3.03 相關性係數的計算方法](#)
- [4. 4.04 相關性≠因果性](#)
- [5. 5.05 相關性分析的侷限性](#)
- [6. 6.06 總結](#)

以下文章來源於數據分析星球，作者數據分析星球

## 1.01 什麼是相關性分析？

當我們面對海量資料時，如何從中提取有價值的資訊？相關性分析是資料分析中非常重要的一環，它可以幫助我們瞭解資料之間的關係，為我們做出更好的決策提供依據。在本篇文章中，我們將結合一個實際的業務場景，來介紹相關性分析的基本概念、步驟和應用，並探討其侷限性和注意事項。

相關性是指兩個或多個變數之間的關係程度。在資料分析中，我們通常使用相關係數來衡量變數之間的相關程度。常見的相關係數包括皮爾森積動差相關係數、斯皮爾曼等級相關係數、切比雪夫相關係數等等。

其中，皮爾森積動差相關係數是最為常見的一種，它可以用來衡量兩個變數之間的線性關係程度。皮爾森積動差相關係數的取值範圍為-1到1，當相關係數接近1時，說明兩個變數之間的正相關性非常強；當相關係數接近-1時，說明兩個變數之間的負相關性非常強；當相關係數接近0時，說明兩個變數之間沒有線性關係。

需要注意的是，相關性並不代表因果關係。兩個變數之間的相關性只是表明它們之間存在某種聯絡或關聯，但並不一定能夠說明其中一個變數的變化是導致另一個變數發生變化的原因。因此，在進行相關性分析時，我們需要同時考慮其他因素，以避免誤判。

## 2.02 相關性分析實際案例

我們以一個銷售業務場景為例來介紹相關性分析的應用。假設某家公司銷售兩種產品：A和B，每月的銷售額和廣告投入如下表所示：

現在，我們想要分析廣告投入與產品銷售額之間的關係，以便更好地制定銷售策略。

首先，我們可以使用皮爾森積動差相關係數來計算廣告投入與產品銷售額之間的相關性。下面是具體的步驟：

1. 計算每個月產品A和產品B的銷售額的平均值和標準差。
2. 計算每個月廣告投入的平均值和標準差。
3. 計算產品A銷售額和廣告投入、產品B銷售額和廣告投入之間的皮爾森積動差相關係數。

根據上述步驟，我們可以得到以下結果：

從上表中可以看出，廣告投入與產品A銷售額、產品B銷售額之間的皮爾森積動差相關係數均非常高，分別為0.981。也就是說，廣告投入與銷售額之間存在非常強的正相關關係。這個結論可以幫助公司制定更好的廣告投入策略，進一步提高銷售額。

除了皮爾森積動差相關係數外，還有其他的相關係數，如斯皮爾曼等級相關係數。在某些情況下，非線性關係可能更為顯著，此時可以使用斯皮爾曼等級相關係數進行分析。

### 3.03 相關性係數的計算方法

計算相關性係數的方法有很多種，下面介紹一些常用的方法和工具。

#### Excel實現

Excel 中的相關性函式為 CORREL，可以用於計算兩個資料系列之間的相關係數。具體使用方法如下：

1. 開啟 Excel 並新建一個工作簿。
2. 在需要計算相關係數的兩個資料系列的單元格中輸入資料。
3. 選中一個空白單元格，輸入 =CORREL(資料系列1, 資料系列2)，按下回車鍵即可計算出兩個資料系列之間的相關係數。

#### SQL實現

在 SQL 中，可以使用 CORR 函式計算相關係數。具體使用方法如下：

1. 開啟 SQL 工具並連線到資料庫。
2. 編寫 SQL 語句，使用 CORR 函式計算兩個資料列之間的相關係數，例如：

#### Python實現

在 Python 中，可以使用 numpy 庫中的 corrcoef 函式來計算相關係數。具體使用方法如下：

這將返回一個 2x2 的陣列，其中第一行第二列和第二行第一列的值就是相關係數。

除了以上介紹的方法，還有一些其他的方法和工具可以用於計算相關係數，例如 MATLAB、R 等。根據實際情況選擇合適的工具和方法，可以快速、準確地計算出相關係數。

### 4.04 相關性≠因果性

雖然相關性分析在資料分析中非常重要，但是我們需要有一些注意事項。

最需要注意的一點是：相關性並不代表因果關係，兩個變數之間的相關性只是表明它們之間存在某種聯絡或關聯，但並不一定能夠說明其中一個變數的變化是導致另一個變數發生變化的原因。因此，在進行相關性分析時，我們需要同時考慮其他因素，以避免誤判。下面我們分別給出一個生活中和資料分析工作中的例子。

#### 生活中的例子

在生活中，有一個經典的例子是冰淇淋銷量和溺水人數之間的相關性。這個例子指出，冰淇淋銷量和溺水人數之間存在正相關關係。也就是說，當冰淇淋銷量增加時，溺水人數也會增加。然而，這並不意味著冰淇淋銷量是導致溺水人數增加的原因。實際上，這個例子中的相關性是由一個更為基礎的因素引起的，即天氣炎熱。當天氣炎熱時，人們更傾向於購買冰淇淋，同時也更傾向於到水中游泳，從而導致了冰淇淋銷量和溺水人數之間的正相關關係。

#### 資料分析工作中的例子

在資料分析工作中，有一個例子是網站流量和使用者購買量之間的相關性。在分析這兩個變數之間的關係時，我們可能會發現它們之間存在正相關關係。也就是說，當網站流量增加時，使用者購買量也會增加。然而，這並不意味著網站流量是導致使用者購買量增加的原因。實際上，這個例子中的相關性是由其他一些因素引起的，比如行銷活動的效果、產品質量、使用者口碑等。因此，在資料分析工作中，我們需要透過更深入的分析，才能確定這兩個變數之間的因果關係，從而制定出更為有效的策略和措施。

## 5. 05 相關性分析的侷限性

儘管相關性分析可以幫助我們理解不同變數之間的關係，但是它也存在一些侷限性，主要表現在：

1. 相關性分析只能衡量線性關係，對於非線性關係，其表現可能不如預期。此時，可以使用其他的相關係數進行分析。
2. 相關性分析只能衡量兩個變數之間的關係，而現實中往往存在多個變數之間的相互作用。在這種情況下，我們需要採用更為複雜的統計模型，如迴歸分析等。
3. 相關性可能是偶然的。在一些情況下，兩個變數之間的相關性可能只是偶然的。例如，在進行大量的資料分析時，有時候會發現兩個變數之間存在很高的相關性，但是這並不代表它們之間存在真正的關係。

## 6. 06 總結

相關性分析是資料分析中非常重要的一環，可以幫助我們瞭解資料之間的關係，為我們做出更好的決策提供依據。在實際應用中，我們需要根據具體問題選擇合適的相關係數進行分析，並注意相關性分析的侷限性和注意事項。

---