frlily / February 25, 2025 11:12AM
“正態分佈”在資料分析中的應用

以下文章來源於數據分析星球 ，作者數據分析星球

1. 01 舉個例子

在生活中，身高是一個常見的連續變數，而且大多數人的身高分佈符合正態分佈。例如，假設我們測量了一個班級中所有學生的身高，並畫出了身高的頻率分佈直方圖。如果這個分佈呈現出鐘形曲線的形狀，那麼這個分佈就可以被認為是正態分佈。在正態分佈中，大多數人的身高會集中在中間，而極端的高或低身高的人數則較少。

正態分佈是統計學中常用的一種分佈類別，它也被稱為高斯分佈或鐘形曲線。正態分佈的特點是具有單峰、對稱、連續和無限可分性等特點。它的機率密度函式具有一個峰值，峰值處的機率最大，並且在峰值兩側逐漸減小，呈現出一條平滑的鐘形曲線。正態分佈在生活中和資料分析工作中都有廣泛的應用。

2. 02 為什麼會出現正態分佈？

正態分佈是一種統計學上的機率分佈模型，它是自然界和社會現象中最常見的分佈之一。從自然界規律的角度來解釋這種現象，我們可以從以下幾個方面進行闡述：

中心極限定理

中心極限定理是統計學中的一個基本定理，它指出當樣本量足夠大時，任何隨機變數的均值分佈將趨近於正態分佈。這個定理可以解釋為，在自然界和社會現象中，許多現象是由許多不同因素的綜合作用而形成的，這些因素的影響是隨機的，而且通常是相互獨立的。因此，隨著資料量的增加，這些隨機因素的影響將趨於平均化，產生一個近似正態分佈的結果。

自然界的複雜性

自然界中的許多生物和物種都具有複雜的生理和行為特徵。例如，身高、體重和壽命等生物學變數通常受到許多基因和環境因素的影響。由於這些因素的影響是隨機的，它們可能會產生一個接近正態分佈的結果。

人類社會的複雜性

人類社會和經濟活動也具有相當的複雜性。例如，收入、財富和教育水平等變數通常受到許多社會、文化和經濟因素的影響。這些因素的影響通常是隨機的，並且可能在不同的群體之間呈現出正態分佈的形式。

所以，正態分佈在自然界和社會現象中非常常見，這是由於許多因素的隨機性和獨立性作用於複雜的生物、自然和社會系統而產生的結果。

3. 03資料分析中正態分佈使用場景

在資料分析工作中，正態分佈是非常重要的概念，因為它可以幫助我們判斷資料是否符合某些假設，以及確定使用哪種統計方法。以下是一些資料分析工作中需要使用正態分佈的場景：

假設檢驗

在假設檢驗中，我們需要假設資料是從一個已知分佈中隨機抽取的。如果我們假設資料來自正態分佈，那麼就需要檢驗資料是否符合正態分佈。許多假設檢驗的方法都基於正態分佈的假設。例如，當我們需要檢驗兩個樣本的平均值是否相等時，我們可以使用t檢驗。但是，t檢驗的前提條件是樣本符合正態分佈。如果資料不符合正態分佈，則需要使用非引數檢驗方法。

迴歸分析

在迴歸分析中，我們通常假設因變數在各自的自變數取值下是正態分佈的。如果資料不符合正態分佈，我們可能需要對資料進行轉換，使其更符合正態分佈。

統計建模

在許多統計建模中，我們需要假設響應變數（例如銷售額）的分佈符合正態分佈。如果響應變數不符合正態分佈，則需要採用其他建模方法，例如廣義線性模型或非引數方法。

控制圖

控制圖是一種品質控制工具，可以幫助我們監控過程是否處於控制狀態。控制圖中的控制限也是基於正態分佈的假設計算出來的。

4. 04 資料分析中正確使用正態分佈

在資料分析中，正確使用正態分佈可以幫助我們做出更準確和可靠的統計推斷。以下是一些使用正態分佈的建議：

正態性檢驗

在使用正態分佈進行假設檢驗或模型構建之前，需要先進行正態性檢驗以確保資料符合正態分佈。

· 繪製直方圖或密度圖：繪製直方圖或密度圖可以幫助我們觀察資料的分佈情況，並判斷是否符合正態分佈。如果資料呈現出鐘形曲線的形狀，那麼它很可能是正態分佈。

· 使用相關工具和技術：在資料分析中，有許多工具和技術可以幫助我們使用正態分佈進行分析，例如正態分佈表、正態機率圖、Q-Q圖等。

· 進行正態性檢驗：進行正態性檢驗可以幫助我們確定資料是否符合正態分佈。在資料分析中，有很多方法可以檢驗資料的正態性，例如Shapiro-Wilk檢驗、Kolmogorov-Smirnov檢驗、Anderson-Darling檢驗等。但需要注意的是，即使正態性檢驗的結果顯示資料不符合正態分佈，也不一定意味著我們不能使用基於正態分佈的方法，因為有些方法對資料分佈的偏離並不敏感。

正態性變換

如果資料不符合正態分佈，我們可以嘗試對資料進行變換，使其更接近於正態分佈。例如，可以嘗試對數變換、平方根變換或Box-Cox變換等。

理解正態分佈的性質

正確理解正態分佈的性質，在進行統計分析時，瞭解正態分佈的性質可以幫助我們更好地理解資料。例如，正態分佈有一個平均值和標準差，這些統計量可以用來描述資料的中心和變異程度。在進行假設檢驗或建模時，我們需要知道正態分佈的均值和標準差的性質，以便進行正確的統計推斷。

正態分佈與抽樣誤差

正確理解正態分佈與抽樣誤差的關係，在資料分析中，我們通常會從樣本中進行推斷整個總體的性質。正態分佈與中心極限定理的關係，可以幫助我們理解樣本大小對抽樣誤差的影響。如果樣本足夠大，即使總體不符合正態分佈，樣本均值的分佈也會趨近於正態分佈。

謹慎使用

雖然正態分佈在許多情況下非常有用，但並不是所有資料都符合正態分佈。在使用正態分佈時，需要注意資料的特徵，以便確定是否適用於該分佈。

5. 05 總結

總之，正態分佈是資料分析中非常重要的概念，它可以幫助我們判斷資料是否符合某些假設，以及確定使用哪種統計方法。在資料分析工作中，我們需要正確理解和使用正態分佈，以避免誤解資料分佈和誤用統計方法。