Vol 06/ Issue 02/ Juin 2024 / pp 170 – 183
ISSN : 2676-234X / EISSN : 2716-9006

Social Empowerment Journal
University Amar Telidji of Laghouat-Algeria

Data review and Analysis

Aicha SOUFRANI
University Amar Telidji of Laghouat, Algeria, e-mail : a.soufrani@lagh-univ.dz
Economics and Management Sciences Laboratory

ORCID: https://orcid.org/0009-0007-2675-7836

Received: 05/02/2024; **Accepted**: 07/04/2024, **Published:** 30/06/2024

Keywords

Mechanisms;
Methods;
Outliers;
Missing value.

Abstract

The studies and Researches have always problems of missing values and outliers, which are often dealt by ignoring or deleting, resulting from it the inaccuracy and validity of the results. So, this study came to refer to both (missing data and outliers) beginning with defining their causes and passing through the mechanisms, then to the most common methods in dealing with this type of data, in order to increase the reliability of the results of studies and research.

Corresponding Author : Soufrani Aicha
DOI 10.34118/sej.v6i2.3929

مجلة التمكين الاجتماعي
جامعة عمار ثليجي الأغواط – الجزائر

المجلد 06 / العـدد 02 / جوان 2024 / ص ص 170 – 183
ISSN : 2676-234X / EISSN : 2716-9006

مراجعة البيانات وتحليلها

عـائشـة صفـراني
جامعة عمار ثليجي الأغواط- الجزائر، البريد الإلكتروني: a.soufrani@lagh-univ.dz
مخبر العلوم الاقتصادية وعلوم التسيير
ORCID: https://orcid.org/0009-0007-2675-7836

تاريخ الاستلام: 2024/02/05 - **تاريخ القبول:** 2024/04/07 - **تاريخ النشر:** 2024/06/30

الكلمات المفتاحية

آليات الفقد؛ طرق المعالجة؛ قيم شاذة؛ قيم مفقودة.

الملخص

لا تخلو البحوث والدراسات من مشكلتي القيم المفقودة والقيم الشاذة، والتي غالبا ما يتم التعامل معها بالتجاهل أو الحذف الأمر الذي يؤدي إلى عدم دقة النتائج وصحتها، لذلك جاءت هذه الدراسة للإشارة لكل منهما بداية بتعريف أسبابهما ومرورا بالآليات (آليات الفقد) وصولا لأهم الطرق الشائعة في التعامل مع هذه البيانات (المفقودة والشاذة)، للعمل على زيادة موثوقية النتائج بالدراسات والأبحاث.

المؤلف المرسل: عـائشـة صفـراني

DOI 10.34118/sej.v6i2.3929

1- مقدمة:

يعد البحث العلمي وسيلتنا للوصول لحل المشكلات واكتشاف الحقائق، والوصول للمعرفة العلمية من خلال إتباع مجموعة من الخطوات والتي تبدأ بتعريف مشكلة البحث وتحديدها وطرح التساؤلات حولها ووضع الفرضيات التي تعد حلول مبدئية لها لاختبارها لاحقا والتأكد من مدى صحتها من خلال جمع البيانات والمعلومات من مصادرها المختلفة وتحليلها ومعالجتها، وصولا لمجموعة من النتائج.

فجمع البيانات واحدة من أهم خطوات البحث العلمي، خاصة إذا تمت معالجتها وتحليلها بالطريقة الصحيحة، ذلك أن الباحث قد يواجه العديد من الحالات والمشكلات التي قد تحول دون الوصول للنتائج المرجوة، إلا أن مراجعة البيانات وتجهيزها وإعدادها لغايات التحليل الإحصائي يحول دونما الوصول لذلك، ولعل في مقدمتها مشكلات وجود القيم الشاذة (Outliers) والبيانات المفقودة (Missing Data). لذلك جاءت هذه الدراسة للبحث بداية في موضوع مراجعة البيانات بالتركيز على تدقيق ومعالجة البيانات المفقودة والقيم الشاذة، ثم الانتقال لطبيعة توزيع البيانات والتي يتحدد من خلالها طبيعة الاختبارات التي يعتمدها الباحث في بحثه (الاختبارات المعلمية واللامعلمية).

ومنه نطرح التساؤل الجوهري التالي: **ماذا نقصد بالبيانات المفقودة والشاذة؟ وكيف يتم التعامل معهما لأغراض إتمام وانجاز البحث؟**

أهداف البحث:

نهدف من وراء هذه الورقة البحثية:

- دراسة القيم المفقودة والقيم الشاذة.
- الكشف عن أسباب فقد البيانات؛ ومنه على الآليات للتعامل معها.
- إظهار أسباب ظهور القيم الشاذة؛ والكشف عن كيفية التعامل مع القيم الشاذة.

أهمية البحث: تتجسد أهمية البحث في الموضوع الذي نتناوله بداية من خلال توضيح مفهوم القيم المفقودة والشاذة وأسبابهما، ومن ثمة الإشارة لأهم الطرق الكفيلة بمعالجة تلك المشكلة التي تحول دونما الوصول لتحليل دقيق للنتائج. مع الإشارة كذلك لأهم الافتراضات في البيانات الخاصة بأهم الاختبارات.

2- جمع البيانات وتحليلها:

قبل التطرق لجمع البيانات لابد من الإشارة لأنواع البيانات التي يحتاجها الباحث ومصادرها، وعموما هناك نوعان من البيانات أولية وثانوية.

2-1- أنواع البيانات:

- **البيانات الثانوية Secondary Data** وهي التي يتم تجميعها في فترات زمنية سابقة ويتم نشرها لأسباب مختلفة قد لا تكون متفقة بدرجة كبيرة مع أهداف الدراسات التي تقوم بها المؤسسات أو الشركات من وقت لآخر وذلك لاختلاف المضمون والنطاق والنتائج لها بالمقارنة مع البيانات التي يتم الحصول عليها من خلال الدراسات الميدانية (مبيضين، 2006، صفحة 29)، وهي البيانات الموجودة

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
https://creativecommons.org/licenses/by-nc/4.0/deed.en

حاليا في المراجع والمصادر (الكتب، المقالات، التقارير الرسمية...) والتي أوجدها باحثين آخرين لأغراض أخرى غير أغراض هذا البحث.

- **البيانات الأولية The Primary Data** وهي التي يبدأ العمل للحصول عليها من خلال تنفيذ مختلف مراحل البحث العلمي. يمكن تجميع هذا النوع من البيانات الأولية إما عن طريق المسوحات الشاملة- إذا كان مجتمع الدراسة صغيرا يمكن التعامل مع كافة مفرداته- وإما عن طريق عينات ممثلة لمجتمع الدراسة من الأفراد أو المؤسسات (مبيضين، 2006، صفحة 29).

كما تعرف البيانات الأولية بأنها البيانات التي قام الباحث بجمعها من مصادرها الأساسية وهي عبارة عن بيانات واقعية وأصلية تعبر عن مشكلة الدراسة، والغرض من جمع البيانات الأولية هو الوصول إلى حل المشكلة البحثية حيث تتضمن البيانات الأولية كلاً من (المقابلة، الاستبيان، الملاحظة) (العفيفي، 2022)، فهي البيانات التي جمعت لأغراض إتمام البحث الحالي ويستخدم الباحث المقابلة والملاحظة والاستبيان للحصول عليها، من خلال تصميم الأداة المناسبة بما يتعلق بمشكلة البحث الحالية التي يعمل على جمع البيانات والمعلومات لحلها.

2-2- مراجعة البيانات وتحليلها:

بعد قيام الباحث بتجميع البيانات بطريقة الاستبانة أو المقابلة أو الملاحظة، تأتي الخطوة التالية وهي عملية تجهيز البيانات وإعدادها لغايات التحليل الإحصائي ليتم الوصول إلى نتائج البحث. ويلجأ معظم الباحثون في الوقت الحاضر إلى استخدام الحاسوب للمساعدة في عملية التحليل نظرا لما يوفره للباحث من توفير للوقت والجهد وسهولة ودقة في استخراج النتائج، ومن أهم البرامج الإحصائية التي تستخدم في هذا المجال برنامج الـ SPSS (عبيدات، أبو نصار، و مبيضين، 2006، صفحة 109)

بعد جمع البيانات تأتي **مرحلة تدقيق** لهذه البيانات تحضيرا لها لعملية التحليل واستخراج النتائج باستخدام الاختبارات الإحصائية الملائمة لكل دراسة حسب المشكلة وأهدافها من خلال عملية وصف البيانات بداية وعرضها في رسومات وجداول وصولا للاختبارات لإثبات صحة أو نفي الفرضيات.

فمن خلال التدقيق للبيانات يقوم الباحث باستبعاد بعض الاستبيانات لأسباب منها:

- عدم دقة بعض الإجابات من قبل المفحوصين وذلك لعدم جديتهم ويظهر ذلك من خلال الإجابات المتناقضة خاصة لما يضع الباحث عبارة وما يناقضها للتأكد من جدية المبحوثين؛
- الإجابات نفسها لكل العبارات؛
- ترك أكثر من نصف الاستبانة بدون إجابة....

وبالتالي لابد من التعامل مع هذه المشكلات بإجراء مراجعة شاملة باستبعاد ما يتم استبعاده، وإهمال ما ليس له علاقة مباشرة بالموضوع واستكمال ما هو ناقص من بيانات بتعويضها (يتجه الكثير من الباحثين لتعويض البيانات الناقصة باستخدام قيمة الوسط وفي كثير من الأحيان النظر في اتجاه أكثر الإجابات

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
https://creativecommons.org/licenses/by-nc/4.0/deed.en

واعتماده)، وهو ما سيتم الإشارة له لاحقا بالإشارة للطرق الشائعة في ذلك بالإضافة للتعامل مع القيم الشاذة- لتحليل البيانات وصولا للنتائج المتوخاة.

3- القيم الشاذة (Outliers):

3-1- تعريف القيم الشاذة:

القيم الشاذة هي مجموعة المشاهدات التي تبعد قيمها بصورة كبيرة جدا عن قيم المشاهدات الأخرى (قويدر و السوالمة، 2017، صفحة 21).

كما تعرف بأنها مفردات أو مشاهدات لها قيم أكبر بكثير أو أقل بكثير من بقية المفردات التي يحتويها المتغير، ويمكن أن تكون هذه القيم الشاذة لمفردات على متغير واحد أو على عدة متغيرات في نفس الوقت (دودين، 2018، صفحة 43).

وهي القيم التي تبتعد كثيرا عن خط الانحدار ويكون حد الخطأ لها كبيرا مقارنة ببقية القيم الطبيعية الأخرى، ويكون لها تأثير كبير على النموذج الخطي ومعلماته. وقد عرفها بارنت(Barnett) بأنها المشاهدة التي تبدو غير منطقية إذ قورنت بسائر البيانات، وعرفها الجبوري بأنها: تلك القيمة التي تكون غير منسجمة مع بقية بيانات المجموعة لمتغير من المتغيرات لظاهرة معينة أو مجموعة من الظواهر، أو أن القيم الشاذة هي القيم التي تأتي من مجتمع يختلف عن مجتمع الدراسة، وعرفها بروس (Bross) بأنها المشاهدة التي تظهر منحرفة بشكل كبير عن سائر مكونات العينة التي وجدت فيها تلك العينة، أما فريمان Freeman فقد عرفها بأنها أي مشاهدة لم تتولد بالطريقة العامة التي ولدت الأغلبية العظمى من مشاهدات البيانات (أبو قديري، 2016).

ومنه نستنتج أن القيم الشاذة هي القيم المتطرفة، أي التي تأخذ قيمة كبيرة جدا أو صغيرة جدا مقارنة بالقيم الأخرى.

3-2- أسباب ظهور القيم الشاذة:

إن أسباب وجود مثل هذه القيم الشاذة في البيانات كثيرة نذكر منها ما يلي (دودين، 2018، صفحة 43):

- وقوع خطأ أثناء إدخال قيمة ما، فمثلا يمكن أن تدخل (50) خطأ بدلا من (5).
- قراءة بيانات مفقودة وكأنها بيانات حقيقية.
- القيمة الشاذة ليست عنصرا في مجتمع الدراسة.
- البيانات نفسها تحتوي على قيم شاذة (بعض أفراد المجتمع مختلفون عن الآخرين بشكل كبير وخصوصا في البيانات التي لا سقف طبيعي لها كالدخل الشهري مثلا).

كما قد تظهر القيم الشاذة لأسباب عدة منها (قاسم و اسماعيل، 2008، الصفحات 71-72):

- إن البيانات تعود إلى توزيعات غير متماثلة أي يكون فيها التواء عال نحو اليمين أو اليسار. ولصياغة نموذج لهذه التوزيعات قدم Green تصنيفا لعوائل توزيعات إحصائية والتي تكون عرضة للقيم الشاذة (Outlier-Prone) وتوزيعات تكون مقاومة للقيم الشاذة (Outlier-Resistant) التوزيعات التي تكون عرضة للقيم الشاذة لها نهايات تؤول إلى الصفر ببطء وهذه التوزيعات تكون عرضة للقيم

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
https://creativecommons.org/licenses/by-nc/4.0/deed.en

الشاذة بصورة مطلقة. والتوزيعات التي نهايتها تؤول إلى الصفر بشكل أسرع من سابقتها وتكون مقاومة للقيم الشاذة بصورة مطلقة.

- تأتي البيانات من نوعين من التوزيعات أحدهما التوزيع الأساسي Basic Distribution والذي يولد مشاهدات جيدة بينما الآخر يسمى التوزيع الملوث Contaminating Distribution والذي يولد قيما شاذة.
- قد تحدث القيم الشاذة نتيجة لأسباب أخرى منها أخطاء يقع فيها الباحث عند تسجيل القياسات، أو نتيجة وجود خلل في جهاز القياسات وخاصة في التجارب المختبرية، أو نتيجة أخطاء في الحسابات مما يؤدي إلى ظهور القيم الشاذة.

2-3- طرق التعامل مع القيم الشاذة في البيانات: وتتمثل في (عواد، 2019، الصفحات 528-529):

- **الحذف:** ويمكن استبعادها من عملية تحليل البيانات إن كانت ذات تأثير كبير، ويرى قاسم واسماعيل والنعيمي أن ظهور القيم المتطرفة في مجموعة البيانات يؤثر بشكل كبير في تحليلها، وفي معظم الأوقات يجب حذف القيم المتطرفة في البيانات، ويؤكد (Al Amri Rahman) على أنه في كثير من الأحيان يتم إزالة القيم الشاذة لتحسين دقة معاملات التقدير، أو يتم استبدالها بقيم أخرى للحصول على تقديرات أدق لمعالم التوزيع.
- **الاحتفاظ:** قد يتم الاحتفاظ بها إن كانت عديمة التأثير أو ذات تأثير قليل في نتائج التحليلات، واقترح أوليوزي (Olewuezi) أن هناك عدد من الاحتمالات التي تستخدم لمعالجة القيم المتطرفة في البيانات ومن ضمنها الاحتفاظ بالقيم المتطرفة لأنه ذا فائدة في نتائج التحليلات الإحصائية وتفسير الظواهر المدروسة.
- **الاستبدال:** حيث يتم معالجتها باستخدام الأساليب الاحصائية المناسبة، ومن أهمها:
  - ✓ طريقة الوسط الحسابي التعويضي (Winsorized Mean): يتم إيجاد وسط حسابي لمجموعة من البيانات تم تقدير القيم المتطرفة فيها عن طريق قيم قريبة منها بدل حذفها.
  - ✓ طريقة الوسط الحسابي المبتور (Trimmed Mean): من طرق علاج القيم المتطرفة في البيانات، وتتسم بالدقة والسهولة وتتم بترتيب القيم في العمود بشكل تصاعدي ثم حذف أصغر قيمة وأكبر قيمة في بيانات العمود، وإيجاد الوسط الحسابي للبيانات المتبقية، ويكون بمثابة القيمة التقديرية للقيم المتطرفة، أي إيجاد وسط مبتور لـ(n-2) من القيم. وافترض توكي (Tukey)أن عدد القيم المراد بترها في الطرفين متساوية، وذلك لصعوبة تحديد موقع القيم المتطرفة، وهذه الطريقة مميزة وتعطي أفضل النتائج عند معالجة القيم المتطرفة.

واستبدال القيم الشاذة بقيمة المتوسط المشذب (Trimmed mean) تتلخص خطواته بترتيب المشاهدات التي تحتوي قيما شاذة تصاعديا أو تنازليا، وتحذف أكبر قيمة وأصغر قيمة، ثم إيجاد الوسط

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
https://creativecommons.org/licenses/by-nc/4.0/deed.en

الحسابي للقيم المتبقية، أي إيجاد الوسط الحسابي المشذب والذي يمثل تقديرا للقيم الشاذة (أبو قديري، 2016)

وتعد طريقة الوسط المبتور Trimmed Mean من الطرق الشائعة في استبدال القيم الشاذة، وتمتاز هذه الطريقة بالدقة والسهولة، وتتم من خلال ترتيب البيانات ترتيبا تصاعديا وحساب قيمة الوسيط، ومن ثم تقدير القيم الشاذة حسب صغرها أو كبرها مقارنة مع بقية البيانات. فإذا كانت القيمة الشاذة أصغر من قيمة الوسيط، يتم حذف أكبر قيمة في البيانات والقيمة الشاذة المراد تقديرها، ومن ثمة إيجاد الوسط الحسابي للقيم المتبقية والذي يعد تقديرا للقيمة الشاذة، وإذا كانت القيمة الشاذة أكبر من قيمة الوسيط، يتم حذف أصغر قيمة في البيانات وحذف القيمة الشاذة المراد تقديرها، وإيجاد الوسط الحسابي للبيانات المتبقية، والذي يعد تقديرا للقيمة الشاذة، وهكذا مع بقية القيم الشاذة (عواد، 2019، صفحة 529).

4- البيانات المفقودة (Missing Data):

من الأمور أو المشاكل المهمة التي يجب التعامل معها جيدا قبل البدء بتحليل النتائج إحصائيا في البحوث والدراسات القيم أو البيانات المفقودة، وتفقد البيانات لأسباب كثيرة منها عدم الدقة في الإجابة أو التطبيق أم عدم الجدية في التعامل مع البحث أو الدراسة العلمية، أو حساسية الموضوع مما يدفع البعض إلى ترك سؤال أو الانسحاب من البحث قبل إكماله وغيرها الكثير من الأسباب، ومهما تكن الأسباب، فقد ينتهي الوضع في الكثير من الأحيان إلى وجود بيانات ناقصة أو مفقودة، وعلى الباحث أن يتعامل مع هذا الوضع (دودين، 2018، صفحة 47).

4-1- تعريف القيم المفقودة وأسبابها:

وتعرف القيم المفقودة Missing Value (صالح، 2019، صفحة 434)

عرفها الرحيل والدرابسة: هي عدم الاستجابة على بعض مفردات مقياس أو اختبار ما من قبل المفحوص، وترك هذه المفردات فارغة دون إجابة.

كما عرفها Graham: بأنها عدم إكمال المفحوص الإجابة عن عبارات المقياس أو الاختبار، وتنشأ هذه المشكلة لعدد من الأسباب؛ مثل عدم استطاعة المفحوص الاستجابة على كل عبارات المقياس بسبب الملل/التعب، أو رفض الإجابة عن سؤال معين، أو رفض المشاركة في اختبار البعدي لدراسة طولية، أو بعض هذه الأسباب معا أو مجتمعة.

وقد أشار ليودلو وأوليري(Ludloz and Oleary) أن سبب وجود القيم المفقودة هو عدم الوصول إلى بعض الفقرات لعدم وجود الوقت الكافي أو لعدم الاهتمام من المستجيب، أيضا أن عدم إجابة المستجيب لبعض الفقرات بغير قصد أو لعدم قدرته على الإجابة يؤدي إلى وجود القيم المفقودة. كما ذكر ميكانيت وآخرون (McKnight et al) ثلاث فئات لأسباب الفقد للبيانات وهي أولا أسباب تعود إلى فئة المستجيبين أنفسهم حول خاصية معينة تتعلق بهم مثلا الدخل الشهري وثانيا منها أسباب تعود لتصميم الدراسة؛ أي أنها

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
https://creativecommons.org/licenses/by-nc/4.0/deed.en

تحتاج لوقت طويل من المستجيب للإجابة عليها وأخيرا أسباب تعود للتفاعل بين المستجيبين وتصميم الدراسة (اللصاصمة، 2016).

وقد ذكر بيو وأندرس (Peugh and Enders)إلى أن كثير من الباحثين يعالجون القيم المفقودة في أبحاثهم بالإهمال. على الرغم من أن هذه القيم قد تكون لها أهمية بتغير نتائج البحث. الأمر الذي يؤدي إلى نتائج غير دقيقة، لذا يجب معالجة القيم المفقودة بطرائق التعويض المناسبة. ويجدر الإشارة إلى أن العلماء صنفوا طرق التعامل مع البيانات المفقودة ومعالجتها إلى طرق قائمة على الحذف (Methods Depend on Deletion) وطرق قائمة على احتساب قيمة تعويضية (Imputation) للقيم المفقودة (اللصاصمة، 2016).

2-4- أنماط القيم المفقودة:

ذكر الزعبي إلى أنه توجد عدة طرق للتعامل مع القيم المفقودة، حيث يساعد معرفة الباحث للنمط (Pattern) التي تظهر عليه القيم المفقودة، وكذلك معرفته لآلية الفقد (Mechanism) بسبب الفقد- على اختيار الطريقة المناسبة للتعامل مع القيم المفقودة.

ميز ليتل وروبين (Little and Rubin) بين ثلاثة أنواع من أنماط فقد القيم هي: النمط الافتراضي أو الاعتباطي (Arbitrary Pattern)، والنمط وحيد المتغير (Univariate Pattern) والنمط الوتيري (Monotone Pattern)، في حين أشار أندرس إلى أن الأدب السابق ميز بين ستة أنواع من أنماط فقد القيم هي (اللصاصمة، 2016):

- **النمط الاعتباطي**: أو النمط العام، في هذا النمط تكون القيم المفقودة منتشرة بشكل عشوائي (دون شكل معين).
- **النمط وحيد المتغير**: ويحدث هذا النمط مثلا عندما تكون هناك فقرة في الاختبار أو الاستبانة لها حساسية عند بعض الأفراد. بمعنى أن القيم المفقودة متعلقة بفقرة واحدة فقط من فقرات المقياس أو الاختبار (متغير واحد)، إذ يوجد عدد من المستجيبين لم يستجيبوا على تلك الفقرة بينما باقي الفقرات أو المتغيرات تحوي بيانات كاملة.
- **النمط الوتيري**: حيث يظهر في هذا النمط أثر الهدر لأفراد العينة إذ يقرر بعض الأفراد الانسحاب من الدراسة بعد المرحلة الأولى ثم يقرر أفراد آخرون الانسحاب بعد المرحلة الثانية وهكذا، فتظهر القيم المفقودة على شكل درج بحيث أن القيم المفقودة تزداد مع ازدياد المرحلة أو مع صعوبة الفقرة.
- **نمط وحدة عدم الاستجابة (Unit Nonresponse Pattern)**: بمعنى أنه لو كان هناك ثلاث متغيرات اثنان منها متوفر بياناتها لجميع المستجيبين، والمتغير الأخر يرفض بعض المستجيبين الاجابة عليه، وغالبا ما يحدث هذا النمط في البحوث المسحية.

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
https://creativecommons.org/licenses/by-nc/4.0/deed.en

- **نمط البيانات المفقودة المخطط لها (The planned Missing Data Pattern):** حيث يتم التخطيط لهذا الفقد من قبل الباحثين عند عملية جمع البيانات وتجهيز أدوات الدراسة ويعتبر هذا النمط ذو فائدة عند جمع بيانات مقياس يتضمن عدد كبير من الفقرات حيث يتم تقسيم فقرات المقياس إلى ثلاثة أجزاء مثلا، وتشكيل ثلاث صور للمقياس كل صورة تحتوي على جزء معين.
- **نمط المتغير الكامن (Latent Variable Pattern):** في هذا النمط تفقد البيانات بسبب متغير كامن ولجميع المستجيبين بالرغم أنه ليس من الضروري عرض المتغير الكامن في نماذج التحليل المستخدمة كمشكلة فقد البيانات حيث تبنى الباحثون خوارزميات لفقد البيانات لتقدير هذه النماذج.

4-3- آليات الفقدان:

باختلاف البيانات غير التامة تختلف الطرائق الإحصائية لتحليل هذه البيانات وأيضا تختلف الآلية التي تؤدي إلى فقدان البيانات وان فهم هذه الآلية وتحديد طبيعتها يساعد كثيرا في اختيار الطريقة المناسبة للتحليل بل يعد المدخل لتشخيص الطريقة التي تقترب نتائجها من الأمثلية للبيانات المدروسة. وتكون علاقة فقدان البيانات لمتغير معين بقيم المتغير نفسه أو لقيم المتغيرات الأخرى وكما يأتي (نعيمي، 2010):

- أن فقدان قيم Xj يكون مستقلا عن قيم المتغيرات الأخرى وعن القيم المفقودة نفسها.
- أن فقدان قيم Xj يعتمد على القيمة نفسها.
- اعتماد القيم المفقودة Xj عن قيم المتغيرات الأخرى في العينة.

إذا كان سبب الفقدان مستقلا عن القيمة المفقودة وعن قيم المتغيرات الأخرى نفسها في العينة عندها يمكن القول أن البيانات تفقد تماما بشكل عشوائي Miss Completely at Random (MCAR) . أما إذا كان سبب الفقدان له علاقة بقيم المتغيرات الأخرى فقط ومستقلا عن القيمة المفقودة ففي مثل هذه يكون فقدان البيانات بشكل عشوائي Missing at Random (MAR) ، وأحيانا يكون سبب الفقدان ناتجا عن القيمة المفقودة نفسها ومستقل عن قيم المتغيرات الأخرى. فالبيانات هنا لا تفقد بشكل عشوائي (Not MAR) عند تحليل هذا النوع من البيانات ويجب أخذ آلية الفقدان بنظر الاعتبار. أما في حالة (MCAR) و(MAR)يمكن أن يهمل التوزيع إليه.

ويمكن التأكد من الفقد العشوائي الكامل للبيانات من خلال مجموعة من الاختبارات، مثل اختبار ليتل (Little :1988) المتوفر في برنامج الرزمة الإحصائية للعلوم الاجتماعية SPSS(اللصاصمة، 2016)

4-4- طرائق معالجة القيم المفقودة:

تتعدد الطرائق التي يمكن من خلالها معالجة القيم المفقودة، ويمكن عرض هذه الطرق كما يلي (الرحيل و الدرابسة، 2014، الصفحات 25-26):

4-4-1- الطرق التي تقوم على الحذف Methods Depends of Deletion: تستخدم هذه الطرق لمعالجة القيم المفقودة، وذلك من أجل إظهار البيانات التي تتضمن القيم المفقودة على شكل بيانات كاملة، ولكن يعاب على هذه الطرق في المعالجة بأنها غالبا ما تعطي نتائج متحيزة وغير فعالة.

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
https://creativecommons.org/licenses/by-nc/4.0/deed.en

2-4-4- الطرق القائمة على احتساب قيمة تعويضية Methods Depends on Imputation: وتقوم هذه الطرق على تقدير قيم معينة وتعويضها بدلا من القيم المفقودة، ومن هذه الطرق:

- حساب قيمة تعويضية واحدة (Single Imputation): ويتم تصنيف الطرق القائمة على احتساب قيمة تعويضية واحدة إلى فئتين هما:

✓ الطرق الصريحة (Explicit Imputation): وفي هذه الطرق يتم استخدام نظام احصائي يمكن الباحث من استبدال المفقودة بقيم مقدرة بطريقتين:

- حساب القيمة التعويضية من خلال المتوسط (Mean Imputation) وفي هذه الحالة يتم حساب القيمة التعويضية للقيم المفقودة بأسلوبين الاول: يتم حساب متوسط العلامات المتوفرة على الفقرة من خلال استجابات المفحوصين عليها، ثم يتم تعويض هذا المتوسط بدلا من جميع القيم المفقودة على هذه الفقرة والثاني: يتم حساب المتوسط الحسابي للمفحوص الواحد من خلال استجاباته على جميع فقرات الاختبار، ثم يتم تعويض المتوسط بدلا من جميع الفقرات المفقودة لهذا المفحوص.
- حساب قيمة تعويضية من خلال الانحدار (Regression Imputation): وتستخدم هذه الطريقة لتقدير القيم التي سيتم تعويضها بدلا من القيم المفقودة، وذلك من خلال تكوين مصفوفة الارتباطات الأساسية للمتغيرات، وكل متغير يتضمن قيما مفقودة، تتم معاملته على انه متغير تابع من خلال معادلة الانحدار التي يتم تكوينها لكل فقرة تتضمن قيما مفقودة، ثم تستخدم المعادلات الناتجة، في الحصول على تقديرات للقيم المفقودة لكل متغير، وبعد ذلك تتم عملية إدخال أو تعويض هذه التقديرات في مجموعة البيانات الناقصة التي تتضمن قيما مفقودة، والقيم المتنبئ بها من معادلة خط الانحدار ، يتم تعويضها بدلا من القيم المفقودة بكل فقرة، وهكذا تكرر هذه العملية لكل فقرة تتضمن قيما مفقودة.

✓ الطرق الضمنية (Implicit Methods): وعند تطبيق هذه الطرق في التعامل مع القيم المفقودة فإنه يتم الاعتماد على أداء الأفراد المفحوصين واستجاباتهم في حساب القيم التي سيتم تعويضها في الفقرات المفقودة. وتشمل هذه الطرق: طريقة حساب القيمة التعويضية بطريقة دالة الاستجابة (Response Function Imputation)، طريقة خوارزمية تعظيم التوقعات (Expectation Maximization Algorithm).

- طريقة حساب قيم تعويضية متعددة Multiple Imputation Method: في هذه الطريقة يتم استبدال كل قيمة مفقودة بمتوسط مجموعة من القيم المختارة عشوائيا، ولذلك ينظر إليها على أنها تقدم قيما تعويضية بأخطاء معيارية غير متحيزة في التحاليل الاحصائية، وهو ما يختلف عن طريقة حساب القيمة التعويضية الواحدة.

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
https://creativecommons.org/licenses/by-nc/4.0/deed.en

- طريقة حساب قيمة تعويضية للوسط المصحح للفقرة: وفي هذه الطريقة يتم تعويض القيم المفقودة للمفحوص، وذلك من خلال استجاباتها واستجابات المفحوصين الآخرين على نفس الاختبار.
- طريقة (الصحيحة جزئيا) (FR) Fractionally Correct Method: وتتعامل هذه الطريقة مع الفقرة المفقودة كأنها صحيحة جزئيا في حال استخدام النموذج الثلاثي المعلم (PL3)، بمعنى أنه عندما يكون عدد الخيارات (Alternatives) للفقرة المفقودة هو (4) خيارات، وتكون العلامة المخصصة للفقرة ذات الاجابة الصحيحة هو علامة واحدة(1)، فإن القيمة التي سيتم تعويضها بدلا من القيمة المفقودة للفقرة، والتي سيتم اعتبارها صحيحة جزئيا وفقا لهذه الطريقة هي (0.25)، وذلك بقسمة العلامة المخصصة للفقرة المفقودة على عدد خياراتها، ثم بعد ذلك يتم تعويض القيم المفقودة لجميع الفقرات المفقودة في الاختبار.
- طريقة حساب القيمة التعويضية من توزيع مشروط Imputing from Conditional Distribution Method: وهذه الطريقة تمزج بين طريقة الانحدار والاختيار العشوائي، وفيها نقوم بتكوين معادلة انحدار لكل فقرة، أو تكوين عدد من المعادلات بطرق مختلفة لنفس الفقرة، ثم يتم اختيار أحد هذه المعادلات عشوائيا، وبوساطتها يتم الحصول على تقدير للقيمة المفقودة.
- طريقة حساب القيمة التعويضية من توزيع غير مشروط Imputing from Unconditional Distribution Method: وبحسب هذه الطريقة يتم احتساب قيمة تعويضية للقيم المفقودة للمفحوص من خلال الاختيار العشوائي لإحدى القيم من بين الاستجابات الموجودة على الفقرة للمفحوصين.
- طرق تقدير قدرات الأفراد ومعالم الفقرات Abilities and Parameters Estimation Methods: ويعد تقدير القدرة للمفحوصين مكونا أساسيا في الاختبارات للمقارنة بينهم، وهناك عدة طرق تستخدم لتقدير قدرة المفحوصين، منها: طريقة الأرجحية العظمى(Maximum Likelihood Estimation)، الطريقة البيزية (Bayesian Method Estimation)، طريقة التقدير الموزونة (Biweight Estimation Method).

5- توزيع البيانات:

بعد جمع ومعالجة البيانات (المفقودة والشاذة)، تأتي مرحلة استخراج النتائج من خلال الاعتماد على الطرق الإحصائية الاستدلالية والتي تصنف إلى الطرق المعلمية واللامعلمية، إلا أن اعتماد أحدها يتطلب الوفاء بافتراضات معينة حول المجتمع الذي تسحب منه العينة، وعلى سبيل المثال التوزيع الطبيعي لذلك سيتم التركيز عليه هنا.

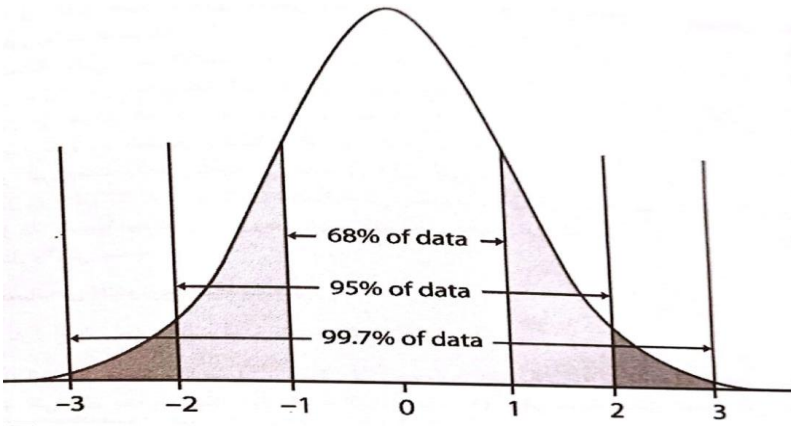
5-1- التوزيع الطبيعي:

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
https://creativecommons.org/licenses/by-nc/4.0/deed.en

إن أكثر التوزيعات تطبيقا هو التوزيع الطبيعي، فعندما تكون مجموعة القيم موزعة بصورة متماثلة حول معدلها، والتوزيع الطبيعي هو توزيع متماثل جرسي الشكل، له قمة واحدة في الوسط وفيه تتساوى قيم الوسط الحسابي والوسيط والمنوال.

وأبرز سمة لهذا التوزيع هي تماثل جانبيه الأيمن والأيسر، ومثل هذه الحالة تعني عدم الانحياز في توزيع القيم، ومركز القيم يمثل القمة، أو التكرارات الأكثر حدوثا. في مجموع القيم الموزعة بصورة متماثلة حول المعدل (التوزيع الطبيعي). تمثل خاصية هذا الشكل تكرار التوزيعات بأعلى كمية في الوسط وأن ذيلي الجرس تمثل الأجزاء الأكثر بعدا عن المعدل، والأقل تكرارا، والأقل احتمالية في الحدوث، أي أن تكرار حدوث قيم المجموعة الموزعة طبيعيا يتناقص تدريجيا في الاتجاهين بعيدا عن المعدل وبشكل متماثل، وتتمثل خصائص هذا التوزيع في (السواعي، 2011، صفحة 129):

- تتراوح الدرجات المعيارية بين (-3 و+3)،
- المنحنى متماثل الجانبين، وذلك لأن مجموع القيم تفوق المعدل يساوي مجموع القيم التي تقل عنه، ولهذا تمثل قيمة المعدل بـ (0). وفي حالة تطابق المعدل مع الوسيط فإن عدد القيم التي تفوق المعدل يساوي عدد القيم التي تقل عنه.
- المنطقة تحت المنحنى معروفة حيث يقع 68.27% بين (-1 و+1)، و95.45% بين (-2 و+2)، و99.73% يقع بين (-3 و+3).



شكل 1. التوزيع الطبيعي

المصدر: (السواعي، 2011، صفحة 130)

ولاختبار التوزيع الطبيعي إحصائيا باستخدام برنامج الـ SPSS، نلجأ لاختبار Shapiro-Wilk في حال العينة أقل من 50 مفردة، و Kolmogorov-Smirnovفي حال العينة أكثر من 50 مفردة.

وعلى العموم فإن عدم تحقق شرط التوزيع الطبيعي للمتغير لا يعتبر مشكلة إذا ما توافر في المتغير المقصود 30 مفردة أو أكثر، ويفسر ذلك بنظرية Central Limit Theorem، والتي تبين أنه إذا اخترنا جميع العينات الممكنة من مجتمع ما، وحسبنا الوسط الحسابي لكل عينة، فإننا سنجد أن توزيع جميع الأوساط

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
https://creativecommons.org/licenses/by-nc/4.0/deed.en

الحسابية لهذه العينات قريب من التوزيع الطبيعي حتى لو لم يكن التوزيع الأصلي للمجتمع قريبا من التوزيع الطبيعي ولكن بشرط أن يكون في كل عينة 30 فردا على الأقل (دودين، 2018، صفحة 59).

2-5- الاحصاءات المعلمية والاحصاءات اللامعلمية:

بعد جمع ووصف البيانات، تأتي مرحلة الاختبار والتحليل وذلك باللجوء لإحدى الطرق الإحصائية المعلمية أو اللامعلمية والتي تساعدنا في ذلك والتي تتوقف كما قلنا سابقا على مجموعة من الافتراضات أهمها التوزيع الطبيعي الذي يحدد طبيعة توجه تحليلنا وفيما يلي عرض مختصر لأهم هذه الطرق والاختبارات.

1-2-5- الطرق المعلمية (Parametric statitic):

الإحصاءات المعلمية هي تلك الطرق التي تتطلب الوفاء بافتراضات معينة حول المجتمع الذي تسحب منه العينة، ومن هذه الافتراضات أن تتخذ المشاهدات في المجتمع شكل التوزيع الطبيعي على سبيل المثال لا الحصر. وقد اشتق مصطلح "معلمية" من مفهوم "معلم" الذي يعني صفة أو خاصية من خصائص مجتمع معين، فلكل قيمة من القيم التي تتعلق بخصائص المجتمع تسمى "معلم Parameter" أما تلك الخصائص المتعلقة بالعينة التي سحبت عشوائيا من المجتمع فتسمى كل منها تقديراً Estimate، أي أن القيمة المستخرجة لأية خاصية في العينة ما هي إلا تقدير لقيمة تلك الخاصية في المجتمع والتي على الأغلب تكون غير معروفة، إذ أنها لو كانت معروفة لانتفت الحاجة إلى حساب تقديرها (السواعي، 2011، صفحة 133).

ومن أشهر الاساليب الإحصائية المعلمية اختبار ت t-test والذي يقوم على افتراض أن العينة المسحوبة من مجتمع إحصائي توزيعه معتدل أو قريب من الاعتدال. بناء على ذلك يستطيع الباحث استخدام توزيع t لإجراء اختبارات الفروض وإيجاد حدود الثقة لمتوسط المجتمع أو الفرق بين متوسطين عندما تكون أحجام العينات صغيرة وتباينات المجتمعات مجهولة (القادر، 2020، صفحة 383).

كما نجد اختبار تحليل التباين ANOVA، الانحدار Regression لقياس الأثر -أثر المتغيرات المستقلة على التابع-، والارتباط Correllation لقياس العلاقة بين المتغيرين ودرجتها....الخ.

2-2-5- الطرق اللامعلمية (Non-parametric statistics):

هي تلك الطرق التي تستخدم في تحليل البيانات واختبار الفرضيات الخاصة بالبيانات الاسمية والرتيبة والتي تكون من النوع المتقطع (المنفصل) عادة ويمكن استخدامها مع البيانات الفترية والنسبية بعد أن يتم تحويلها إلى بيانات اسمية أو رتبية، أو تلك البيانات التي لم تف ببعض الافتراضات الأساسية Assumptions مثل تجانس التباين في حالة تحليل التباين ANOVAواختبار ت t-test. هذا فضلا عن أن هذه الطرق لا تتقيد بالشروط الواجب توافرها لاستخدام الإحصاءات المعلمية، حيث إنها تستخدم في الحالات التي لا يكون فيها التوزيع النظري للمجتمع الأصلي الذي اختيرت منه العينة معروفا، أو في حالة عدم إمكانية الوفاء بافتراض أن التوزيع النظري للمجتمع طبيعيا، ولا بضرورة أن يكون اختيار العينة من ذلك المجتمع عشوائيا (السواعي، 2011، صفحة 133).

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
https://creativecommons.org/licenses/by-nc/4.0/deed.en

ومن أكثر الاختبارات اللامعلمية استخداماً، اختبار مان-وتني، اختبار ويلكوكسون، اختبار كروكسال-واليس، وارتباط سبيرمان.

ويعد الإحصاء المعلمي أدق وأكثر كفاءة من الإحصاء اللامعلمي، كما أنه أكثر حساسية لخصائص البيانات التي يتم جمعها، كما أنه يوفر فرصة ضئيلة لحدوث الخطأ من النوع الأول والخطأ من النوع الثاني، ويؤخذ على الاختبارات المعلمية بأنها أكثر صعوبة عند حسابها، بالإضافة إلى محدودية نوع البيانات التي يمكن اختبارها بواسطة تلك الاختبارات وتستغرق وقتا وجهدا في تطبيقها ولكن مع توفر البرامج الإحصائية مثل SPSS أمكن التغلب على هذه السلبيات (القادر، 2020، صفحة 384).

6- الخاتمة:

تطرقنا في بحثنا إلى مشكلة تواجه أغلب الباحثين، عند قيامهم بجمع بياناتهم وهي عدم اكتمال الإجابات من جهة ووجود قيم متطرفة من جهة أخرى مما يؤدي إلى تقديرات متحيزة أو أقل كفاءة، ويحول دون وصول الاختبارات لمعلومات مهمة ودقيقة، وقد توصلنا لمجموعة من النتائج والتوصيات نوردها كما يلي:

النتائج:

- أن حل مشكلة البيانات المفقودة والشاذة من البداية تقود الباحث لدقة التقدير ولموثوقية النتائج.
- أن التعرف على آلية الفقدان للبيانات تتيح للباحث اختيار الأسلوب والطريقة الملائمة لمعالجة الفقد.
- ضرورة الاعتماد على الطريقة المناسبة لمعالجة القيم المفقودة والشاذة.
- أن الاختبارات الاحصائية (المعلمية واللامعلمية) تتوقف على طبيعة توزيع البيانات.

كما نقترح ونوصي بـ:

- الكشف عن القيم الشاذة في البيانات قبل التحليل لأن نتائج التحليل قد تختلف بوجود القيم الشاذة من عدمها.
- معالجة القيم المفقودة والقيم الشاذة قبل إجراء أي اختبار.
- إجراء دراسة مقارنة بين طرق المعالجة لتحديد أفضل الطرق في معالجة سواء القيم المفقودة أو القيم الشاذة.

6- قائمة المراجع:

أسوان، محمد طيب نعيمي. (2010). معالجة البيانات غير التامة وتقديرها بطريقة المكونات الرئيسية. *المجلة العراقية للعلوم الاحصائية*، الصفحات 327-338.

جميل، فرهود أبو قديري. (2016). استخدام البواقي والقيم الشاذة للكشف عن انتهاكات افتراضات تحليل الانحدار الخطي البسيط. الأردن: جامعة مؤتة.

حمزة، محمد دودين. (2018). *التحليل الاحصائي المتقدم للبيانات باستخدام SPSS*. عمان. الأردن: دار المسيرة.

خالد، محمد السواعي. (2011). *مدخل إلى تحليل البيانات باستخدام SPSS* (الإصدار ط1). اربد-الأردن: عالم الكتب الحديث.

راتب، صايل الخضر الرحيل. رياض، أحمد صالح الدرابسة. (حزيران، 2014). أثر طريقي التعامل مع القيم المفقودة، وطريقة تقدير القدرة على دقة تقدير معالم الفقرات والأفراد. *المجلة الدولية التربوية المتخصصة، 3*(6)، الصفحات 23-47.

علي، محمد العرسان بني عواد. (2019). القيم الشاذة في أداء الطلبة على اختباري القدرات والتحصيل وأثر أسلوب التعامل معها في نتائج التحليلات الاحصائية. *مجلة البحث العلمي في التربية* (20)، الصفحات 525-542.

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
https://creativecommons.org/licenses/by-nc/4.0/deed.en

عمر، قاسم قويدر. يوسف، محمد السوالمة. (2017). القيم الشاذة في أداء الطلبة الأردنيين على اختبار (TIMSS) في الرياضيات والعلوم وأثر أسلوب التعامل معها في نتائج التحليلات الاحصائية. *مجلة جامعة القدس المفتوحة للأبحاث والدراسات التربوية والنفسية* (20)، الصفحات 18-33.

عمران، اسماعيل اللصاصمة. (2016). أثر نسبة القيم المفقودة وطريقة معالجتها في دقة تقدير معالم معادلة الانحدار البسيط. الأردن: *جامعة مؤتة*.

فيصل، أحمد العبد القادر. (2020). حجم تأثير الاختبارات الاحصائية المعلمية واللامعلمية المستخدمة في رسائل الماجستير بكلية التربية بجامعة الملك سعود. *المجلة العلمية لكلية التربية-جامعة اسيوط، 26*(4)، الصفحات 376-411.

محمد، عبيدات. محمد، أبو نصار. عقلة، مبيضين. (2006). *منهجية البحث العلمي القواعد والمراحل والتطبيقات* (الإصدار ط2). عمان، الأردن: دار وائل.

محمد، نذير اسماعيل قاسم. يونس، حازم اسماعيل. (2008). الكشف عن القيم الشاذة بأسلوب بيز باستخدام معاينة جبس. *المجلة العراقية للعلوم الاحصائية* (14)، الصفحات 68-88.

نوال، جبار صالح. (2019). المقارنة بين تقديرات معالم انموذج راش للبيانات الكاملة والمفقودة باختلاف طرق معالجة البيانات المفقودة. *مجلة البحوث التربوية والنفسية، 16* (63)، 429-466.

طارق، العفيفي. (2022)، الفرق بين البيانات الأولية والثانوية، تم الاسترداد: 2023/07/19. من: https://drasah.com/Description.aspx?id=5703

- Arabic references in English:

Aswan, M. T. N. (2010). Incomplete data processing and estimation by the main components method. Iraqi Journal of statistical sciences, pp. 327-338.

Jamil, F. A. (2016). The use of residuals and anomalous values to detect violations of the assumptions of simple linear regression analysis. Jordan: Mutah University.

Hamza, M. D. (2018). Advanced statistical analysis of data using SPSS. Oman. Jordan: Dar Al-Masirah.

Khalid, M. A. (2011). An introduction to data analysis using SPSS (version i1). Irbid-Jordan: the modern world of books.

Ratib, S. A.A. Riad, A. S. A. (June, 2014). The impact of the two methods of dealing with missing values, and the method of estimating the ability to accurately estimate the parameters of paragraphs and individuals. International Journal of specialized pedagogy, 3 (6), pp. 23-47.

Ali, M. A. B. (2019). Abnormal values in students ' performance on the aptitude and achievement tests and the impact of the method of dealing with them on the results of statistical analyses. Journal of scientific research in education (20), pp. 525-542.

Omar, Q. K. Youssef, M. A. (2017). Abnormal values in the performance of Jordanian students on the TIMSS test in mathematics and science and the impact of the method of dealing with them on the results of statistical analyses. Journal of al-Quds Open University for educational and psychological research and studies (20), pp. 18-33.

Omran, I. A. (2016). The effect of the ratio of missing values and the method of their processing on the accuracy of estimating the parameters of the simple regression equation. Jordan: Mutah University.

Faisal, A. A. (2020). The size of the impact of the parametric and non-parametric statistical tests used in the master's theses at the Faculty of Education, King Saud University. Scientific journal of the Faculty of Education-Assiut University, 26(4), pp. 376-411.

Mohammad, A. Mohammad, A. Okla, M. (2006). Scientific research methodology rules, stages and applications (version i2). Amman, Jordan: Wael House.

Mohammad, N. I. Q. Youness, H. I. (2008). Detection of anomalous values in the biz style using a gypsum preview. Iraqi Journal of statistical sciences (14), pp. 68-88.

Nawal, J. S. (2019). Comparison of estimates of the parameters of the Rasch model for complete and missing data by different methods of processing missing data. Journal of educational and psychological research, 16 (63), 429-466.

Tarek, A. (2022), the difference between primary and secondary data, retrieved: 19/07/2023. From: https://drasah.com/Description.aspx?id=5703

***Citation**: Soufrani, A. **Data review and Analysis**. Social Empowerment Journal. 2024; 6(2): pp. 170-183. https://doi.org/10.34118/sej.v6i2.3929*

***Publisher's Note:** SEJ stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.*

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
https://creativecommons.org/licenses/by-nc/4.0/deed.en