القسم الثالث
المعالجة الأولية للبيانات

يتناول هذا القسم عرضاً للخطوة الثالثة من خطوات تكوين المؤشر المُركَّبَ، وهى خطوة المعالجة الأولية للبيانات، حيث يعرض الجزء الأول من هذا القسم تقدير القيم المفقودة بالبيانات، بينما يعرض الجزء الثانى التأكد من مدى جودة البيانات الأساسية من خلال دراسة مدى اتفاق هذه البيانات مع معايير جودة البيانات العالمية.

٣. ١ تقدير البيانات المفقودة (Imputation):

تعتبر البيانات المفقودة من أكثر المعوقات التى تحد من متانة وجودة المؤشر المُركَّبَ، لذلك فقد كان من الضرورى استعراض بعض الطرق التى تُستخدم لمعالجة وتقدير تلك البيانات، إلا أنه قبل استعراض هذه الطرق كان لابد من سرد أنواع البيانات المفقودة، حيث تختلف طرق التقدير تبعاً لاختلاف نوع البيانات المفقودة.

٣. ١. ١ أنواع البيانات المفقودة:

يمكن تقسيم البيانات المفقودة إلى ثلاثة أنواع أساسية: بيانات مفقودة بشكل عشوائى تماماً (Missing completely at random)، بيانات مفقودة بشكل عشوائى (Missing at random)، وبيانات مفقودة بشكل غير عشوائى (Not missing at random)، وفيما يلى عرض لتلك الأنواع بشىء من التفصيل:

(أ) بيانات مفقودة بطريقة عشوائية تماماً Missing completely at random (MCAR):

فى هذا النوع من فقد البيانات نجد أن البيانات المفقودة لا تعتمد على المتغيِّر نفسه – الذى يحتوى على هذه القيم المفقودة – أو أى متغير آخر فى قاعدة البيانات، فعلى سبيل المثال تعتبر القيم المفقودة من متغير الدخل من نوعية البيانات المفقودة بطريقة عشوائية تماماً فى حالة كون القيم المفقودة غير مرتبطة بالدخل أو بأيٍ من المتغيرات الأخرى فى قاعدة البيانات. بمعنى أنه إذا ما كان الأشخاص الذين لم يقرُّوا بقيم دخولهم لهم – فى المتوسط – نفس قيم دخول الأشخاص الذين أقرُّوا بالدخل، أو إذا

كان كل متغير من المتغيرات الأخرى الموجودة فى قاعدة البيانات هو نفسه — فى المتوسط — لكلٍ من الأشخاص الذين أقرُّوا بقيمة الدخل والأشخاص الذين لم يقرُّوا به.

(ب) بيانات مفقودة بطريقة عشوائية Missing at random (MAR):

نجد أن البيانات المفقودة فى هذا النوع لا تعتمد على المتغير محل الدراسة، إلا أنها ترتبط بمتغيرات أخرى فى قاعدة البيانات، فعلى سبيل المثال قد تعتمد نسبة البيانات المفقودة للدخل على الحالة الاجتماعية، إلا أنه فى داخل كل فئة من فئات الحالة الاجتماعية تكون نسبة البيانات المفقودة غير مرتبطة بقيمة الدخل، وهذا ما يعنى أن القيم المفقودة لا تعتمد على متغير الدخل إلا أنها تعتمد على الحالة الاجتماعية للأفراد، وكذلك من حالات فقد البيانات بطريقة عشوائية هى حالة البيانات المفقودة نتيجة لتصميم الإستمارة Missing data by design، مثل حالة عدم انطباق السؤال على المجيب (not applicable) فمثلاً إذا كانت إجابة سؤال معين بنعم فهذا يتبعه عدم الإجابة عن سؤال آخر بالإستمارة.

(جـ) بيانات مفقودة بطريقة غير عشوائية Not missing at random (NMAR):

فى هذه الحالة تعتمد البيانات المفقودة على القيم نفسها، فعلى سبيل المثال، الأفراد ذوى الدخول المرتفعة غالباً لا يذكرون دخولهم. وغالباً لا يوجد أساس معين للحكم على ما إذا كانت البيانات مفقودة بطريقة عشوائية أم بطريقة غير عشوائية، إلا أن معظم طرق تقدير البيانات المفقودة تتطلب أن يكون السبب عشوائياً (MCAR, MAR)، وإذا كان هناك اتجاهات تشير إلى وجود نمط غير عشوائى لفقد البيانات، فإنه يجب توضيح هذا النمط ونمذجته بدقة، إلا أن هذه العملية تتسم بالصعوبة والتعقيد.

٣. ١. ٢ أساليب تقدير البيانات المفقودة:

بعد أن تم التعرُّف على الأنواع الثلاثة لفقد البيانات، يتم الانتقال لعرض الطرق الأساسية التى يمكن استخدامها لمعالجة البيانات المفقودة، حيث تتمثل هذه الطرق فى الآتى:

(أ) حذف المفردة (Case deletion):

تعتمد هذه الطريقة على حذف المشاهدة – التى تحتوى على قيمة مفقودة لأحد المتغيرات – نهائياً من التحليلات، ولذلك فهى تسمى أيضا تحليل الحالة الكاملة (complete case analysis) حيث أنها تعتمد ببساطة على حذف الحالات غير الكاملة من التحليل، إلا أنه من عيوب هذه الطريقة أنها

تتجاهل احتمال وجود اختلافات بين العينات الكاملة وغير الكاملة، كما أنها تتطلب فقد البيانات بطريقة عشوائية تماماً، حيث أنها تؤدى إلى تقديرات غير متحيزة فقط إذا كانت المشاهدات المحذوفة تعتبر عينة جزئية عشوائية من العينة الأصلية. هذا بالإضافة إلى أن الأخطاء المعيارية سوف تكون أكثر فى العينة غير الكاملة نظراً لاستخدام معلومات أقل، ووفقاً للتجربة يمكن القول أن القاعدة العامة التى يمكن الاعتماد عليها فى هذه الطريقة هى عدم حذف أى من الحالات إذا كان المتغير لديه أكثر من ٥٪ من القيم المفقودة.

(ب) التقدير الفردى (Single imputation):

تعتمد هذه الطريقة على تقدير البيانات المفقودة من خلال عمل توزيع تنبؤى (Predictive distribution) لهذه البيانات المفقودة، ويجب أن يُولد هذا التوزيع التنبؤى عن طريق توظيف البيانات من خلال تصميم إما نماذج ظاهرة (Explicit modeling)، أو نماذج ضمنية (Implicit modeling).

تعتمد النماذج الضمنية على نظام حسابى ذو افتراضات ضمنية يجب اختبار كونها مناسبة وجيدة للغرض محل الاهتمام أم لا، ومن أمثلتها:

- **تقدير هوت دك (Hot deck imputation):**

حيث يتم تقدير البيانات المفقودة ببيانات مسحوبة من وحدات مشابهة، على سبيل المثال: القيم المفقودة للدخل لأحد الأفراد يمكن التعويض عنها بقيم الدخل لأفراد لهم نفس الخصائص.

- **التعويض (Substitution):**

تعتمد هذه الطريقة على إحلال الوحدات غير المستجيبة بوحدات لم يتم اختيارها فى العينة، فمثلاً إذا لم يتاح الاتصال بأحد الأسر، فإننا نقوم باختيار أسرة تنتمى إلى نفس الكتلة السكنية لم تكن مختارة فى العينة.

- **تقدير كولد دك (Cold deck imputation):**

حيث يتم استبدال البيان الناقص ببيان من مصدر آخر، كأخذ البيان من إصدار سابق لنفس الاستقصاء.

أما بالنسبة للنماذج الظاهرة، فيعتمد التقدير فيها على نماذج إحصائية حيث تكون الافتراضات واضحة وظاهرة، ومن أمثلة هذه النماذج:

- **التقدير باستخدام المتوسط الحسابى/ الوسيط / المنوال الغير شرطى (Unconditional mean/median/mode imputation):**

وتعتمد هذه الطريقة على استبدال البيانات المفقودة بقيمة المتوسط (mean) – أو الوسيط (median) أو المنوال (mode) – المقدر باستخدام المشاهدات الغير مفقودة.

- **تقدير البيانات عن طريق تحليل الانحدار (Regression imputation):**

تعتمد هذه الطريقة على التعويض عن القيم المفقودة بالقيم المُتنبأ بها من تحليل الانحدار، على أن يكون المتغير التابع فى تحليل الانحدار هو المؤشر الفرعى الذى لديه قيم مفقودة، ويكون المتغير المستقل (المتغيرات المستقلة) هو المؤشر الفرعى (المؤشرات الفرعية) الذى يعكس علاقة قوية مع المتغير التابع، أى توجد درجة عالية من الارتباط بين المتغير المستقل (المتغيرات المستقلة) والمتغير التابع، ومن الجدير بالذكر هنا أنه يجب أن يتم عمل الانحدار باستخدام المشاهدات الكاملة، أى التى لا يوجد بها قيم مفقودة للمتغير محل الاهتمام.

- **التقدير باستخدام القيمة العظمى المتوقَعة (Expected Maximization Imputation):**

تعتمد هذه الطريقة على التعويض عن القيم المفقودة من خلال مقدرات يتم الحصول عليها من خلال عملية تقدير متكررة (Iterative process)، حيث يتم أولاً التنبؤ بالقيم المفقودة بناءً على مقدرات أولية (Initial estimates) لقيم معالم النموذج، ثم تستخدم هذه التنبؤات بعد ذلك لتعديل قيم المعالم، ويتم تكرار عملية التنبؤ باستخدام معالم النموذج المعدلة حتى تقترب المعالم من تقديرات الإمكان الأكبر (Maximum-likelihood estimates)، وبذلك فإن تقدير البيانات المفقودة باستخدام هذه الطريقة يتكون من جزأين: الأول هو تقدير معالم النموذج عن طريق الإمكان الأكبر (Maximum Likelihood) بفرض أنة لا توجد بيانات ناقصة، والثانى هو حساب القيمة المتوقَّعة للبيان المفقود بواسطة المعالم المقدرة.

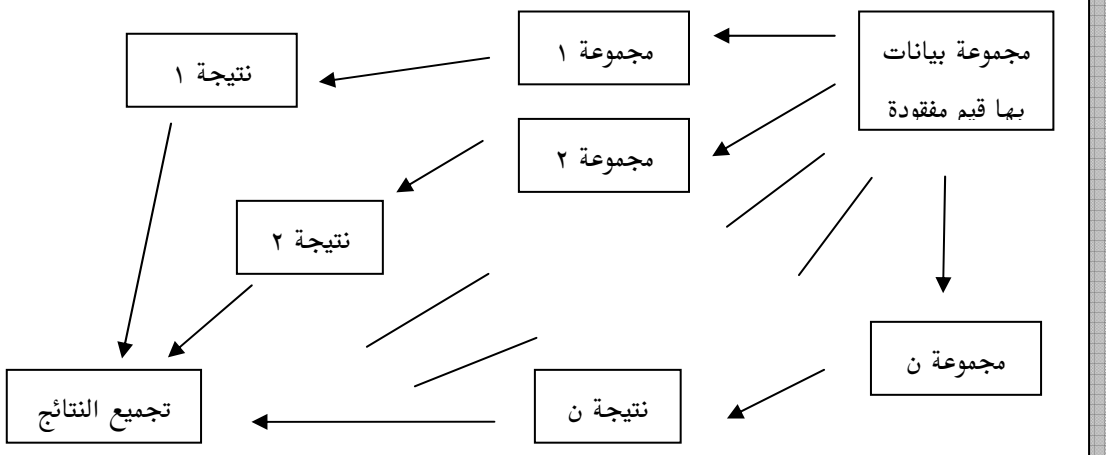
(جـ) التقدير المتعدد للبيانات المفقودة (Multiple imputation):

تعتمد هذه الطريقة على تقدير البيانات المفقودة (ن) مرة، بهدف الحصول على (ن) مجموعة كاملة من البيانات، بحيث أنه فى كل مجموعة من البيانات يتم تقدير المعلمات محل الاهتمام، وكذلك تقدير أخطائها المعيارية، ثم يتم حساب متوسط التقديرات (وسط حسابى أو وسيط) لجميع المجموعات

(ن مجموعة)، وكذلك التباين داخل المجموعات، والتباين بين المجموعات. والشكل التالى يوضح المنطق من وراء التقدير المتعدد للبيانات المفقودة.

شكل رقم (٣)

المنطق من وراء التقدير المتعدد للبيانات المفقودة



ومن الجدير بالذكر هنا أنه يمكن استخدام أى طريقة من طرق التقدير السابقة لتقدير البيانات فى كل مجموعة من المجموعات.

٣. ٢ جودة البيانات الأساسية:

يعتبر الأساس الذى يُبنّى عليه المؤشر المُركَّبَ هو البيانات الأساسية للمؤشرات الفرعية، لذلك فيجب أن يكون الأساس جيد حتى يكون البناء جيد، وحتى نستطيع الاعتماد على المؤشر المُركَّبَ فيما بعد. وحيث إن جودة المؤشر المُركَّبَ الكلية تعتمد على جودة البيانات المبدئية المستخدمة فى بناءه، فإن هذا يستدعى التأكد من جودة البيانات الأساسية حتى يتم اختيارها بشكل يعظم الجودة الكلية للنتائج النهائية.

٣. ٢. ١ معايير الجودة:

البيانات الأساسية المستخدمة فى تكوين المؤشر المُركَّبَ لابد أن تُحقق معايير جودة البيانات الأساسية التى تتمثل فى:

- نفعية البيانات **(Relevance):**

يتم تقييم نفعية البيانات الأساسية المكونة للمؤشر المُرَكَّبَ مع الأخذ فى الاعتبار الهدف النهائى من المؤشر، بحيث يتم تقييم واختيار البيانات الأساسية بحذر لضمان أن هذا المعيار تم تغطيته بشكل مناسب. فمثلاً بافتراض عدم التوافر الفعلى للبيانات، ففى أغلب الأحيان يتم استخدام سلاسل بديلة، ولكن فى هذه الحالة لابد أن تتوافر بعض الأدلة عن علاقة هذه السلاسل البديلة بسلاسل الهدف.

- الدقة **(Accuracy)**:

يعكس هذا المعيار مدى دقة وصف البيانات للظاهرة التى صممت من أجل قياسها، كما يوضح مدى القرب بين القيمة المقدرة والقيمة المجهولة، وتعتبر دقة البيانات الأساسية من المعايير الهامة جداً للجودة، وهذا ما يستدعى أن يتم الحصول على البيانات الأساسية للمؤشر المُرَكَّبَ بمستوى عالى من الدقة، إلا أنه من الناحية العملية لا يوجد مقياس للدقة، ولكنه قد يمكن قياس مستوى الدقة بالبيانات الأساسية، من خلال قياس الخطأ فى المقدرات التى تعتمد على عينات المسوح، ويعتبر أهم مصادر الأخطاء فى هذه المقدرات هو خطأ التغطية، وخطأ المعاينة، وخطأ عدم الإستجابة.

- حداثة البيانات **(Timeliness):**

يتمثل هذا المعيار فى الحصول على البيانات فى الوقت المناسب ووفقاً لتواريخ مسبقة، كما يعكس الفترة ما بين المرحلة البحثية أو نهاية الفترة البحثية والوقت الذى تكون فيه البيانات متاحة، ويعتبر هذا المعيار من الأبعاد الهامة جداً لتقليل الاحتياج إلى تقدير البيانات المفقودة ومراجعة البيانات المنشورة مسبقاً لتنقيحها، ومن الجدير بالذكر فى هذا المعيار هو أن البيانات التى تغطى مجالات مختلفة تكون متاحة فى أوقات مختلفة خلال الزمن، بمعنى أنه قد تكون بيانات كل مؤشر من المؤشرات الفردية متاحة فى وقت مختلف عن المؤشرات الفردية الأخرى، ولذلك فيجب أن يكون هناك تناسب بين توقيتات الحصول على البيانات المكونة للمؤشر المُرَكَّبَ.

- إمكانية الوصول إلى البيانات والحصول عليها **(Accessibility):**

يشير هذا المعيار إلى مدى سهولة الحصول على البيانات، وهذا يتضمن سهولة الوصول إلى البيانات، بالإضافة إلى وضوح طريقة عرض تلك البيانات، ونلاحظ أن سهولة أو صعوبة الوصول

إلى البيانات يؤثر على التكلفة الإجمالية لإنتاج وتجديد المؤشر المُرَكَّبَ خلال الزمن، كذلك يستطيع هذا المعيار التأثير على مصداقية المؤشر المُرَكَّبَ إذا كان هناك صعوبة فى الوصول إلى البيانات، حيث أنه قد يؤثر على المعايير الأخرى، فقد تؤثر صعوبة الوصول إلى البيانات على حداثتها حيث يتم الوصول إلى البيانات فى توقيت غير مناسب، وقد تؤثر صعوبة الوصول للبيان على دقة البيانات التى يتم الحصول عليها، ونلاحظ أنه فى ظل التطورات التكنولوجية والقدرة على الوصول إلى قواعد البيانات المصدرة من جهات مختلفة، أصبح من السهل الوصول إلى الكثير من البيانات، وكذلك إيجاد العلاقات بين المصادر المختلفة للبيانات، وبالتالى الحصول على أكبر قدر ممكن من المنفعية للبيانات، ولكن من الجدير بالذكر هنا أنه لابد عند اختيار مصدر البيانات أن لا يتم تفضيل المصدر السهل الوصول إليه فقط، ولكن يجب النظر فى معايير الجودة الأخرى.

- إمكانية التفسير **(Interpretability):**

يعكس هذا المعيار مدى السهولة التى يستطيع بها المستخدم تفسير البيانات، وذلك ما يتم عن طريق توافر المعلومات الإضافية لتفسير واستخدام البيانات بشكل ملائم، وتشتمل هذه المعلومات عادةً علي المفاهيم الأساسية والمتغيرات والتصنيفات المستخدمة ومنهجية جمع البيانات، ونظراً لاتساع مدى البيانات المستخدمة لبناء المؤشر المُرَكَّبَ، وكذلك الصعوبات المتعلقة بالإجراءات التجميعية المطلوبة،فإن ذلك يحتاج إلى تفسير كامل، ويعتبر هذا المعيار من المعايير الهامة للجودة، حيث أن توافر التعريفات والتصنيفات المستخدمة لإنتاج البيانات يكون ضرورى لتقييم المقارنات بين البيانات عبر الزمن وبين الدول المختلفة.

- التماسك **(Coherence):**

يقيس هذا المعيار درجة سهولة مقارنة البيانات باختلاف الزمان أو المكان، أى تماسك البيانات عبر الزمن وبين الدول المختلفة، ويشير التماسك عبر الزمن ضمنياً إلى أن البيانات مستندة على مفاهيم وتعريفات ومنهجيات عامة ومشتركة عبر الزمن، حيث يشير عدم التماسك عبر الزمن إلى التغيرات التى تحدث فى السلاسل الزمنية نتيجة للتغيُّر فى المفاهيم أو التعريفات أو المنهجيات المستخدمة. أما بالنسبة للوجه الثانى للتماسك، فهو التماسك عبر الدول البلدان، وهو ما يشير ضمنياً إلى أنه من بلد إلى بلد يظل هناك اشتراك فى المفاهيم والتعريفات والتصنيفات والمنهجيات. وتظهر أهمية الارتباط والاتساق المنطقى فى أن المسوح الميدانية الإحصائية قد يتم تطبيقها على

مجتمعات مختلفة وفى أوقات مختلفة، لذلك فيجب أن تكون الإحصاءات الناتجة عنها مترابطة، ويمكن المقارنة بينها حتى تكون بيانات المؤشرات الفرعية متماسكة للدول المختلفة والأزمنة المختلفة.