# مقارنة بين التعويض الاحادي والاستبعاد الثنائي للقيم المفقودة

د.اشرف ادريس سعيد محمد

أستاذ مساعد – قسم الاحصاء والدراسات السكانية
جامعة القضارف.

**المستخلص :**

تهدف هذة الدراسة الى تسيط الضوء علي طرق معالجة البيانات المفقودة وكذلك المشاكل والانعكاسات المترتبة عليها والمقارنة بين طريقة التعويض الاحادي والاستبعاد الثنائي لتقدير القيم المفقودة مع مراعاة اختلاف حجم العينة ، حيث تم استخدام المنهج الوصفي والاستدلالى للوصول الى نتائج الدراسة باستخدام برنامج (SPSS- 20) وتوصلت الدراسة الى أن إستخدام طريقة التعويض بقيمة تعويضية واحدة أكثر فعالية من إستخدام طريقه الإستبعاد الثنائي للقيم المفقودة في حالة زيادة جم العينة.

**الكلمات المفتاحية:** البيانات المفقودة، التعويض الاحادي، التعويض الاحادي، الاستبعاد الثنائي ، حجم العينة.

# Comparison Between Single Imputation And Binary Exclusion For Missing Data

**Ashraf Idress Saeed Mohammed**

**Abstract**

This study aims to investigate the methods of missing Data processing, and also the problems and their reflections. In addition to comparison between single imputation and binary exclusion for missing data estimation in case of different sample size. The researchers used the descriptive and analytical method to draw results by using spss20. The study revealed that using method of single imputation with one value was more effective than using binary exclusion method in case of increasing sample size.

**Keywords: Missing data, Single imputation, Binary exclusion, Statistic product and service solution.**

## المقدمة :

تعتبر البيانات المفقودة من المشاكل البحثية الشائعة أو المتكررة وهي تعني أن يتم فقدان جزء من بيانات العينة لأي سبب وهي البيانات المفقودة من متغير او أكثر والتي تحدث غالبا نتيجة جمع البياتات بطريقة غير صحيحة او إرتكاب أخطاء عند إدخال البيانات مما يؤثر سلبا علي عملية جمع البيانات لذلك يجب أن يكون الباحث علي دراية كافية بالتعامل مع البيانات المفقودة وتكون البداية بأن يحدد الباحث نوع البيانات المفقودة الخاصة به كما يحدد الطريقة المناسبة لمعالجتها في التحليل الإحصائي , وتعد البياتات المفقودة معضلة لأن الأسالب الإحصائية تفترض معلومات كاملة عن جميع المتغيرات المدرجة في التحليل ويمكن لعدد قليل نسبيا من البيانات المفقودة علي بعض المتغيرات يقلص بشكل كبير في حجم العينة نتيجة لذلك تتأثر دقة فترات الثقة وتضعف القوة الإحصائية وتكون معلمة التقدير متحيزة.

## مشكلة البحث:

تعد البيانات المفقودة مشكلة كبيرة للباحث وعدم معالجتها بشكل مناسب قد يسبب للباحث المشكلات التالية (تقليل حجم العينة و عدم تقدير التباين بشكل صحيح و الوصول الى نتائج متحيزة ) .

## أهمية البحث :

تكمن اهمية البحث فى علاج المشاكل التى تواجه الباحثين فى كيفية معالجة البيانات المفقودة وتحليل البيانات بطريقة صحيحة .

## فروض البحث:

1. البيانات المفقودة تؤدي الى الوصول الى النتائج متحيزة في حالة تغيير حجم العينة
2. هنالك علاقة بين البيانات المفقودة وحجم العينة.
3. طريقه التعويض الاحادي افضل من طريقة الأستبعاد الثنائى للقيم المفقودة.

## أهداف البحث:

يهدف البحث الى دراسة المقارنة بين طريقة التعويض الاحادي والاستبعاد الثنائي لمعالجة البيانات المفقودة، وكذلك تسليط الضوء علي المشاكل والانعكاسات المترتبة علي ذلك.

**منهجية البحث :**

إعتمد البحث على المنهج الوصفي و التحليلى باستخدام البرنامج الإحصائى SPSS 20 (المنتج الاحصائي وحلول الخدمة ) .

الجانب النظري :

**أنواع البيانات المفقودة :**

البيانات المفقودة هي القيم المفقودة من متغير أو أكثر والتي تحدث غالباً نتيجة لجمع البيانات بطريقة غير صحيحة أو إرتكاب أخطاء عند إدخال البيانات ويقود الباحث في النهاية الى بناء استنتاجات غير سليمة . ويمكن تقسيم البيانات المفقودة الى الاتي :

I.بيانات مفقودة بصورة عشوائية تامة .

II.بيانات مفقودة بصورة عشوائية .

III.بيانات مفقودة بصورة غير عشوائية . [1]

**كيف يتم التعامل مع البيانات المفقودة :**

هناك عدة طرق للتعامل مع البيانات المفقودة ومنها :

**طريقة الحذف :**

**ويندرج تحتها طريقتين مختلفتين :**

**I.حذف الحالة بالكامل**: تتم استبعاد الإستبيان أو ألإختبار عن فقد أي معلومة منه اغلب الباحثين يستعملون هذه الطريقة لسهولتها رغم المحاذر المرتبطة عليها مثل انخفاض حجم العينة وألتاثير علي قوة الإختبار وقد يقع الباحث بسبب ذلك في أخطاء من النوع الأول أو الثاني , إعطاء نتائج غير حقيقية (تحيز النتائج ) لأن العينة أصبحت متحيزة غير ممثلة للمجتمع [1]

**II.الحذف المزدوج :**

يتم هنا حذف الحالات المفقودة فقط للمتغير الواحد عند حساب علاقته مع متغير اخر وليس حذف كامل للحالة كما في الطريقة السابقة مثلاً إذا كان الباحث يدرس العلاقة بين الدخل الشهرى والمستوى التعليمي فإنه يستبعدهما معاً عند فقد البيانات في إحداهما مع المحافظة علي بقية المتغيرات كالجنس والعمر والذكاء ومن أثار هذه الطريقة هو اختلاف حجم العينة من متغير لأخر من ما يؤثر سلباً علي صدق النتائج [1]

**طريقة التعويض :**

ويندرج تحتها كذلك طريقتين مختلفتين هما :

التعويض بقيم متعددة : ويندرج تحتها الأساليب التالية :

I.Expectation Maximization وإختصارها EMALGORITHM

**تعتمد هذه الطريقة علي خطوتين :**

a.ألاولى توقع القيمة المفقودة

b.الثانية تعظيم الإحتمالية

ويحتاج هذا الإسلوب الى إحصائية كبيرة لدي الباحث ويمكن عملها علي برنامج SPSS

II. Multiple Imputation Methods: إختصارهاMI Methods

تقوم على فكرة تكوين أكثر من قيمة للبيانات المفقودة ثم إستخدام المتوسط لتلك القيم كبديل مناسب.تعتبرأفضل الطرق لأنها لا تعتمد على البيانات المفقودة من نوعMCR أو MARبل هي صالحة لكل الأحوال وبإمكان عملها مع برنامجSPSS. [1]

**التعويض بقيمة واحدة : ويندرج تحتها الأسالب التالة :**

1- Mean Imputation:يتم إستبدال الدرجة المفقودة بالمتوسط أو الوسيط للمتغير نفسه

مثالا إذا كان لدينا 50 شخص ذكروا أعمارهم بينما أمتنع 5أشخاص عن ذلك فيتم حساب متوسط أعمار ال50 شخص ثم يعطي الناتج لكل من رفض إعطاء عمره.

و الإشكالة في هذا الأسلوب هو إستبدال البيانات المفقودة برقم ثابت يؤثر سلبيا على التباين لأفراد العينة. [2]

2 - Hot – dec Imputation:-يتم إستبدال الدرجة المفقودة من خلال البحث عن شخص متقارب معه في مختلف الخصائص الديموغرافية كالمستوى التعليمي أوالعمر ...الخ .

الإشكالة في هذا الأسلوب هو التأثير السلبي علي متوسط المجموعة وتباينها بسبب القيمة المضافة[1].

**الطريقة الرابعة:**

**التعويض بمتوسط الفرد :**

إذا كان لدينا إستبيان مكون من 100 فقرات مثلا وأجاب أحد أفراد العينة على 80 فقرات فقط فيتم إيجاد القيم المفقودة من خلال متوسط إجابات الفرد (يتم القسمةعلى 80 وليس على 100)وأخيرا هنا بعض النقاط الهامة والتي ينبغي على كل باحث التركيز عليها عند ((الكتابة)) عن البيانات المفقودة في إجراءت الدراسة :-

a.الكتابة عن الإجراءات التي أتبعها الباحث لتقليل فقد البيانات

b.بيان حجم العينة المستهدفة وحجم العينة الفعلية

c.بيان نسبة البيانات المفقودة

d.توضيح الة توزيع البيانات المفقودة على المتغيرات والحالات

e.بيان الطريقة الإحصائية المستخدمة لتحديد نوع فقد البيانات [1]

f.بيان الطريقة المستخدمة للتعامل مع البيانات المفقودة مع توضيح أسباب إستخدامها

الاستبعاد الثنائي للبيانات المفقودة: Binary Exclusion of missing Data

في هذة الطريقة نقوم بإجاد الارتباطات لجميع المتغيرات التي تحتوي علي بيانات مفقودة وهنا تم الاعتماد علي متغيرين إثنين (X1 , X2) لكل عينة بحيث يتم حساب معامل ارتباط بيرسون حيث يمكن أن

يعتمد كل ارتباط علي عدد مختلف من الحالات ومن ثم تعويض قيمة الارتباط الناتجة بدل القيم المفقودة ومن ثم اجراء اختبار الوسط للعينتين الجديدتين بعد التعويض للمقارنة [2]

## الجانب التطبيقي :

### 1- توليد البيانات :

لقد تم توليد هذه البيانات عن طريق الألة الحاسبة الرقمية بواسطة EXP وأستخدمت لمعرفة طريقة تعويض القيم المفقودة بطريقتي الإستبعاد الثنائي للقيم المفقودة وطريقة التعويض بقيمة تعويضية واحدة ثم إختيار ثلاثة عينات عشوائية لكل متغيرين :

تم إختيار العينات على النحو التالى:

1-حجم العينة الأولى n=100

2-حجم العينة الثانية n=150

3-حجم العينة الثالثة n=200

**2-الإستبعاد الثنائي للقيمة المفقودة :-BinaryExclusion of missing values**

قمنا بإيجاد الإرتباطات لجميع المتغيرات قيد الدراسة التي تحتوي على بيانات مفقودة من العينة الاولي (X1 يحتوي علي 100 رقم)و كذلك (X2 يحتوي علي 100 رقم) حيث تحتوي كلعينة علي 5% من القيم المفقودة ثم يستخدم كل إرتباط بين المتغيرين للتعويض عن كل البيانات المفقودة في هذين المتغيرين .أجرينا الأختبار للعينة الأولى x1,x2 وكان قيم الإرتباط (36) وتم تعويضها عن كل قيمة مفقودة في x1,x2.

**اولا:** قمنا بإجراء اختبار الوسط للعينتين قبل الحذف وكانت النتائج كالاتي :

$X_1 = 432 , 5900$

$X_2 = 449 , 1300$

**ثانياً :** قمنا بحذف (5%) من البيانات بطريقة عشوائية ومن ثم ايجاد الارتباطات كما في الجدول التالى :

**الجدول (1) يوضح الأرتباط بي المتغيرين :**

| الارتباطات | | | |
|---|---|---|---|
| | | X1 (100) | X2 (100) |
| X1 | معامل ارتباط بيرسون | **1** | **036.-** |
| | القيمة الاحتمالية | | **734.** |
| | العدد | **95** | **90** |
| X2 | معامل ارتباط بيرسون | **036.-** | **1** |
| | القيمة الاحتمالية | **734.** | |
| | العدد | **90** | **94** |

ثم أجرينا إختبار الوسط الحسابي ل $x_1,x_2$ بعد تعويض القيم المفقودة وكانت كالأتي:

$X_1 = 432, 3800$

$X_2 = 449,1300$

ومن هنا نتأكد من صحة تعويض قيم البيانات المفقودة بطريقة الإستبعاد الثنائي للقيم المفقودة ، حيث اظهرت النتائج عدم وجود اختلاف معنوي (sig =0.002) بين متوسطات العينتين قبل وبعد الحذف (X1) و(X2) .

## 3-التعويض بقيمة تعويضية واحدة :

عند زيادة حجم العينة من 100 إلى 150 للتحقيق من الفرضية أجرينا الإختبار للعينة الثانية (100) رقم لكل عينة($x_1$ , $x_2$) وتم إيجاد الوسط الحسابي $x_1$ وكانت 453,2421 وتم تعويض الوسط الحسابي للقيم المفقودة للعينة $x_1$ وثم أجرينا إختبار الوسط الحسابي للعينة بعد تعويض القيم المفقودة للتأكد من صحة التعويض وكانت =453,2300 ،ومن ثم ذيادة جم العينة الى (150) رقم وأجرينا نفس الإختيار للعينة $x_2$ وكان 475,5000 بالتقريب ≈ 476 وثم تعويض الوسط الحسابي في القيم المفقودة للعينة $x_2$ وثم أجرينا إختبار الوسط الحسابي للعينة بعد التعويض للتأكد من صحة التعويض وكان 475,5300 بالتقريب ≈ 476 من الملاحظ هنا عند إختلاف حجم العينة لم يتغير المتوسط .

**الإستبعاد الثنائي للقيم المفقودة:**

الجدول(2) يوضح الأرتباط بين المتغيرين.

| Correlations | | | |
|---|---|---|---|
| | | (x1(150 | (x2(150 |
| x1 | Pearson Correlation | **1** | **060.** |
| | (Sig. (2-tailed | | **482.** |
| | N | **145** | **140** |
| x2 | Pearson Correlation | **060.** | **1** |
| | (Sig. (2-tailed | **482.** | |
| | N | **140** | **144** |

أجرينا الإرتباط بين العينة الجديدة وكانت ,60 وثم حسبنا الوسط الحسابي للعينة بعد تعويض الإرتباط بين القيم المفقودة وكان المتوسط الحسابي للعينة $x_1$= 462,3800 والوسط الحسابي للعينة $x_2$ = 500,4333 وكان الوسط الحسابي للعينة الأصلية قبل التعويض لـــ $x_1$ = 476,2552 وللعينة $x_2$ = 518,7847 وهذا يبن أن إختلاف حجم العينة بالذيادة تؤثر على تعويض القيم المفقودة في حالة الاستبعاد الثنائي .

## 4- التعويض بقيمة تعويضية واحدة:

أجرينا تجربة أخرى وزدنا حجم العينة من 150 إلى 200 وأجرينا فيها نفس الإختبارات السابقة للتأكد من النتائج وكانالوسط الحسابي بالنسبة ل$x_1$ = 476,2552 وللعمود $x_2$ = 518,784 وثم تعويض

الأوساط الحسابية مكان القيم المفقودة وثم إجراء الوسط الحسابي للعينة بعد التعويض وكان للعمود $x_1$ = 476,246 وللعمود $x_2$ = 518,7533 وهذا يبين أن طريقة التعويض بالوسط الحسابي أفضل من طريقة التعويض بالإرتباط .

وللتأكد من ذلك :

**نختبر طريقة الإستبعاد الثنائي للقيم المفقودة :-**

الجدول (3) يوضح الأرتباط بين المتغيرين

| Correlations | | | |
|---|---|---|---|
| | | x1 (200) | x2 (200) |
| x1 | Pearson Correlation | **1** | **012.** |
| | (Sig. (2-tailed | | **868.** |
| | N | **195** | **190** |
| x2 | Pearson Correlation | **012.** | **1** |
| | (Sig. (2-tailed | **868.** | |
| | N | **190** | **194** |

أجرينا نفس الإختبارات السابقة وكانت قيمة الإرتباط بين $x_1$و $x_2$ = 12 وتم حساب الوسط الحسابي للعينة بعد التعويض بالإرتباط وكان الوسط للعمود الأول $x_1$ = 479,0450 وللعمود $x_2$ = 503,1000 وكان الوسط الحسابي للبيانات الأصلية للعمود $x_1$ = 491,0205 وللعمود $x_2$ = 518,2887 وهذا يتبين أن إختلاف العينة أدى إلى إختلاف قيم الاوساط بعد التعويض بطريقة الإستبعاد الثنائي .

## 5- طريقة التعويض بقيمة تعويضية واحدة :

أجرينا نفس الإختبارات السابقة وكان الوسط الحسابي للبيانات الأصلية للعمود $x_1$ = 491,0205 وللعمود $x_2$ = 518,2887 وأجرينا الوسط الحسابي للبيانات بعد التعويض بالوسط وكانت كالأتي الوسط الحسابي للعمود $x_1$ = 491,0200 والوسط الحسابي للعمود $x_2$ = 518,2800 وهذا يثبت أن طريقة التعويض بالوسط الحسابي أفضل من طريقة الإستبعاد الثنائي بالإرتباط.

## الخاتمة :

ومع نهاية هذا البحث نجد ان التعويض عن البيانات المفقودة من الامور الهامة جدا لدي الباحثين ومحللي البيانات اذ انها الركيزة الاساسية لتحسين جودة النتائج المتحصل عليها ، لذلك كان لابد من تسليط الضوء علي هذا الجانب ، معالجة هذة المشكلة تتطلب مجهود كبير يتمثل في كيفية الحصول علي بيانات بها فقد في البايانات وصعوبة معرفة ما اذا كانت البيانات التي تم التعويض عنها تتجانس مع البيانات الموجودة ام لا ، لذلك يمكن توليد بيانات تكون معلومة لدي الباحث ومن ثم حذف جزء بسيط منها يمثل مثلا 2% واجراعملية التعويض بالطريقتين ومن ثم مقارنة البيانات الاصلية لمعرفة ما اذا كانت تتطابق مع البيانات التي تم التعويض عنها .ومن اهم النتائج التي تم التوصل اليها ان استدام طريقة التعويض الاحادي اكثر دقة وفاعلية من استخدام طريقة الاستبعاد الثنائي .

**النتائج :**

أ. أظهرت النتائج أن قيم التعويض بقيمه تعويضية واحدة للبيانات المفقودة صحيحة لان الوسط الحسابي للقيم المفقودة قبل التعويض يساوى الوسط الحسابي للقيم المفقودة بعد التعويض وكانت الأوساط الحسابية للبيانات المفقودة قبل التعويض ل$x_1$:453.5900 وكانت بعد التعويض :453.3800 وأن جميع قيم التعويض بالأستبعاد الثنائي للقيم المفقودة (إستخدام الإرتباط ) هي طريقة صحيحه فقط في العينة الأصلية أما إذا تم تغيير حجم العينه تتغير النتيجة ,والوسط الحسابي للبيانات الأصلية لايساوي الوسط الحسابي للبيانات بعد التعويض .أي إذا زاد حجم العينة في طريقة الإستبعاد الثنائي تتغير النتيجة أي يؤثر حجم العينة علي التعويض بالإستبعاد الثنائي للقيم المفقودة .

ب. أظهرت النتائج أنه حتي بعد تغيير حجم العينة لطريقة التعويض بأستخدام قيمة تعويضية واحدة لاتتغير النتيجة ,أي ان تغيير حجم العينة لايؤثر علي قيمة التعويض باستخدام قيمة تعويضية واحدة .

ت. أن إستخدام طريقة التعويض بقيمة تعويضية واحدة أكثر فعالة من إستخدام طريقه الإستبعاد الثنائي للقيم المفقودة .

**التوصيات :**

أ. إستخدام طريقة التعويض بإستخدام قيمة تعويضية واحدة (التعويض الأحادي) لأنها تصلح حتي في حالة تغيير حجم العينة قيد الدراسة .

ب. لاينصح بإستخدام التعويض عن القيم المفقودة بطريقة الأستبعاد الثنائي للقيم المفقودة لأنها تتأثر بحجم العينه .ولاتصلح إذا تم تغيير حجم العينة قيد الدراسة .

ت. إجراء دراسات تختبر طرائق أخرى في معالجة القيم المفقودة للبيانات لبيانات مقاييس متعددة وولرتبة إستجابة غير ثنائية .

## المراجع والمصادر :


(1) سامى الشهري، البيانات المفقودة ،انواعها والاساليب الاحصائية للتعامل معها،الانترنت (https://ar=lang?1188818526682976257/thread/com.rattibha)، 2019، (2-9).

(2) حسان ابو حسان، الاساليب الاحصائية في معالجة القيم المفقودة في التعداد والمسوحات، (المعهد العربي للتدريب والبحوث الاحصائية ورشة عمل ضمن التعليم عن بعد- فلسطين) (2020)،4-41 .

(3) الحيالي، علي درب كسار. استخدام معايير الدقة التنبؤية في تحديد الطريقة المثلى في تقدير القيمة المفقودة: بيانات البحوث الزراعية أنموذجًا. مجلة العلوم الزراعية العراقية، (2013) ،44(4)، 509-517.

(4) الرحيل، راتب صايل والدرابسة، رياض أحمد. أثر طريقتي التعامل مع القيم المفقودة، وطريقة تقدير القدرة على دقة تقدير معالم الفقرات والأفراد. المجلة الدولية التربوية المتخصصة،(2014) 3(6)، 23-47.

(5) ضعضع، هبة وطومان، منار وطيفور، مصطفى. أثر حجم العينة وطرائق التقدير في دقة تقدير معالم نموذج راش. مجلة جامعة جرش. (2020)، 21(1)، 131-170.

(6) Beale, E. M., & Little, R. J. Missing Values in Multivariate Analysis, Journal of the Royal Statistical Society, (1975) , Series B, 37, 129145-.

(7) Acock AC. Working With Missing Values. Journal of Marriage and Family [Internet]. 2005 Nov 1 [cited 2018 Oct 8];67(4):1012–28. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.17413737.2005.00191-.x

(8) Bennett DA. How can I deal with missing data in my study? Australian and New Zealand Journal of Public Health [Internet]. 2001 Oct 1 [cited 2018 Oct 8];25(5):464–9. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467842-X.2001.tb00294.x

(9) KJ, Carlin JB. Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normalz Imputation. Am J Epidemiol [Internet]. 2010 Mar 1 [cited 2018 Oct 8];171(5):624–32. Available from: https://academic.oup.com/aje/article/171137388/624/5/