# اختبار (Chi-Square) للاستقلالية والتجانس

مقدمة

البيانات التصنيفية Categorical data هي البيانات التي تصنف الحالات في الدراسة إلى عدد من المستويات وتكون قيمة المتغير لحالة ما هي قيمة من عدد من القيم المحددة مسبقا. فعلى سبيل المثال يمكن تصنيف رضاء عميل عن خدمة معين بأنه (كامل، نوعا ما، غير راضي) أو يمكننا تصنيف مصدر السيارات في المملكة إلى (أمريكا، أوروبا، شرق آسيا) أو يمكننا تصنيف تجاوب مريض لدواء بأنه (قوي، ضعيف) وهكذا. لذلك فإن المتغير موضع القياس سيأخذ أحد القيم المحدد ولا يمكن أن يأخذ قيمة أخرى إلا إذا قمنا بتعديل التصنيف ليتلاءم مع القيم الجديدة.

**جداول الاقتران**

عندما يتعلق الأمر بتحليل العلاقة بين متغيرين تصنيفيين (حقليين)، Categorical Variables، فانه يتم التعامل في الأصل مع جدول اقتران، Contingency Table، يتكون من أعمدة وصفوف. تمثل الصفوف المختلفة في جداول الاقتران حقول المتغير التصنيفي الأول بينما تمثل الأعمدة حقول المتغير التصنيفي الثاني، و يتم تعبئة خانات الجدول الداخلية بالتكرارات للمشاهدات التي تقع في الصف المحدد والعمود المحدد. فمثلاً، لدراسة علاقة الحالة الاجتماعية بمستوى التعليم يمكن تلخيص قراءات العينة العشوائية المسحوبة من مجتمع ما والمصنفة حسب المتغيرين التصنيفيين السابقين وذلك من خلال تفريغ بيانات العينة العشوائية في خانات جدول الاقتران. افترض أن X تمثل المتغير الأول وان Y تمثل المتغير الثاني، كذلك افترض أن عدد حقول المتغير X مساوي لــ R بينما عدد حقول المتغير الآخر مساوي لــ C. في هذه الحالة يتم إنشاء جدول اقتران للمتغيرين X و Y والذي سيأخذ الحجم R×C (تقرأ ار في سي). والتي تعطي مؤشر واضح إلى وجود I من الصفوف و J من الأعمدة.

وبافتراض أن التجربة ابتدأت من خلال سحب عينة عشوائية من المجتمع بحجم $N$ وتم بعد ذلك تصنيف وحدات العينة حسب المتغيرين الحقليين، فان الأرقام التي في الخانات المكونة للجدول يطلق عليها تكرارات مشاهدة وسيتم الرمز لها بالرمز O نسبة إلى كلمة Observed. وتمثل المجاميع الهامشية للصفوف والأعمدة تكرارات يتم التعرف عليها بعد تفريغ العينة العشوائية في جدول الاقتران للمتغيرين الحقليين. كذلك يشير الدليل في التكرارات

المشاهدة إلى كل من رقم الصف ورقم العمود على التوالي. فالتكرار المشاهد $O_{ij}$ يشير إلى التكرار للخانة الموجودة في الصف $i$ والعمود $j$. حيث

$$i = 1, 2, \cdots\cdots, r \ \ \& \ \ j = 1, 2, \cdots\cdots, c$$

كذلك يتم إيجاد المجاميع الهامشية عن طريق المعادلات التالية:

$$N_{i\,.} = \sum_{j=1}^{c} O_{ij} \quad \forall \ i = 1,2,.......,r$$

$$N_{.\,j} = \sum_{i=1}^{r} O_{ij} \quad \forall \ j = 1,2,.......,c$$

$$N = \sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij}$$

**مكونات جداول الاقتران**

| | Y | | | | | |
|---|---|---|---|---|---|---|
| | **c** | **...** | **3** | **2** | **1** | **X** |
| $N_{1.}$ | $O_{1c}$ | ... | $O_{13}$ | $O_{12}$ | $O_{11}$ | **1** |
| $N_{2.}$ | $O_{2c}$ | ... | $O_{23}$ | $O_{22}$ | $O_{21}$ | **2** |
| $N_{3.}$ | $O_{3c}$ | ... | $O_{33}$ | $O_{32}$ | $O_{31}$ | **3** |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| $N_{r.}$ | $O_{rc}$ | ... | $O_{r3}$ | $O_{r2}$ | $O_{r1}$ | **r** |
| **N** | $N_{.c}$ | **...** | $N_{.3}$ | $N_{.2}$ | $N_{.1}$ | |

### أسلوب المعاينة

يمثل أسلوب المعاينة المستخدم في الحصول على التكرارات المشاهدة في جداول الاقتران عامل رئيسي في تحديد التوزيعات النسبية أو الاحتمالية للمتغيرات التصنيفية. وبلا شك فإن لأسلوب المعاينة دور رئيسي ومهم جدا في العملية الإحصائية المطبقة على تحليل البيانات التصنيفية. ويتم التفريق بين حالتين أساسيتين وهما:

### *حجم العينة الكلي عشوائي*

عندما يتم سحب عينة عشوائية من $n$ من القراءات أو المشاهدات، ويتم توزيعها على خانات جدول اقتران لمتغيرين تصنيفيين، فان المعاينة هنا يمكن وصفها بالمعاينة العشوائية التامة. وكنتيجة فان مجاميع الصفوف ومجاميع الأعمدة تمثل هنا أرقام عشوائية يتم الحصول عليها بعد إجراء التجربة وتفريغ وحدات العينة العشوائية على خانات جدول الاقتران.

### *حجم العينة للمتغير المستقل عشوائي*

عند تحديد مجاميع الصفوف أو مجاميع الأعمدة قبل إجراء التجربة وسحب العينة العشوائية، فإن المجاميع المحددة مسبقا هي مجاميع ثابتة (Fixed) غير عشوائية، والتي بالطبع ترتبط دائما بالمتغير المستقل (سواء كان المتغير التصنيفي الواقع في الصفوف أو الأعمدة). ويفترض هنا وجود تأثير للمتغير المستقل على المتغير التابع (المتغير التصنيفي الآخر). لذا فان الهدف الأساسي في عملية الاستدلال الإحصائي في هذه الحالة يتمثل بدراسة معنوية تأثير المتغير المستقل على المتغير التابع.

### *اختبار (Chi-Square) للاستقلالية*

يطلق على المتغيرين التصنيفيين صفة الاستقلال عندما يكون الاحتمال المشترك للمتغيرين هو حاصل ضرب الاحتمال الحدي لكلا المتغيرين التصنيفيين. بيد أن اختبار الاستقلال لمتغيرين تصنيفيين يرتبط ارتباط قوي ومباشر، كما سبق الإشارة إليه، بأسلوب المعاينة المتبع في تكوين جدول الاقتران. فعندما يتم تثبيت المجموع الكلي لحالات الدارسة، فإن المجاميع الهامشية للمتغيرين التصنيفيين عشوائية وغير ثابتة. ويكون التساؤل المطروح غالب حول مدى استقلال المتغيرين التصنيفيين. لذا فانه لا يوجد في هذه الحالة متغير مستقل، حيث يمكن اعتبار كلا المتغيرين التصنيفيين متغيرات ذات تأثير تبادلي.

*فرضيات يجب توافرها بالعينة لإجراء الاختبار*

1. البيانات المستخدمة في الدارسة بيانات وصفية، وتعكس هذه الفرضية أن البيانات التي في الخلايا تمثل تكرارات في خلايا متنافية، أي أنه لا يمكن وضع مشاهدة في أكثر من مستوى من مستويات التصنيف.
2. عينة الدراسة عينة عشوائية تتكون من $n$ مشاهدة مستقلة.
3. يجب أن تكون التكرارات المتوقعة أكبر من 5، ويمكن التساهل في هذه الفرضية.

وعند توافر هذه الشروط فإنه يمكن صياغة فرضية العدم والفرضية البديلة على النحو التالي:

$H_0$: المتغيران التصنيفيين مستقلان

$H_a$: الفرضية البديلة: المتغيران التصنيفيين غير مستقلان.

ويمكن صياغة الفرضيتين رياضياً على النحو التالي:

$$H_0 : o_{ij} = \varepsilon_{ij}$$

$H_a : o_{ij} \neq \varepsilon_{ij}$ (لخلية واحدة على الأقل)

ويستخدم اختبار (Chi-Square) لاختبار هذه الفرضية حيث يتم مقارنة التكرارات المتوقعة والتي يتم حسابها بناء على صحة فرضية العدم بالتكرارات المشاهدة. وتستخدم الصيغة الرياضية التالية لحساب إحصائية (Chi-Square).

$$\chi^2 = \sum_{j}^{n_i} \sum_{i=1}^{k} \left[ \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \right]$$

حيث وتمثل $o_{ij}$ التكرارات المشاهدة في حين $e_{ij}$ تمثل التكرارات المتوقعة.

### *اختبار (Chi-Square) للتجانس*

عندما يكون كل من حجم العينة الكلي ($n$) والمجاميع الهامشية لأحد المتغيرين التصنيفيين محددة مسبقا قبل سحب العينة وإجراء التجربة، فان التساؤل يكمن في معرفة مدى تطابق التوزيع لحالات الدراسة للمتغير التابع لكل حقل أو تصنيف للمتغير المستقل. في هذه الحالة تكون المجاميع الهامشية للمتغير المستقل ثابتة ومحددة مسبقا بينما تكون المجاميع الهامشية للمتغير التابع عشوائية تنتج بعد عملية تفريغ البيانات في جدول الاقتران. ويستخدم اختبار $\chi^2$ لجودة التوافق أو التجانس في بحث فرضية اختلاف التوزيع التكراري في الخلايا لكل مستوى من مستويات المتغير التابع عبر مستويات المتغير المستقل. فعند أخذ عينة حجمها $n_i$ من مجتمع ويتم تصنيف هذه العينة إلى $k$ مستوى وذلك بوضع المشاهدات من العينة في جدول تقاطعي بحيث يتكون كل صف من $k$ خلية، وتمثل كل خلية في كل صف من الجدول أحد مستويات التصنيف للمتغير التابع، فإن اختبار $\chi^2$ للتجانس يبحث في ما إذا كان هناك فروق بين التكرارات النسبية المشاهدة من $k$ خلية وبين لتكرارات المتوقعة أو النظرية والتي يتم تحديدها بناء على حاصل ضرب $n_i$ في (نسبة التكرار الإجمالي لكل مستوى من

مستويات المتغير التابع إلى العدد الكلي للمشاهدات). وعند وجود اختلاف بين التكرارات المشاهدة والتكرارات المتوقعة في خلية واحدة على الأقل، فإن الباحث يستنتج أن هناك إمكانية كبيرة في وجود تأثير للمتغير التابع على المتغير المستقل، أو بعبارة أخرى فإن التوزيع النسبي للحالات عبر مستويات المتغير المستقل غير متجانسة على الأقل لأحد مستويات المتغير التابع.

فرضية العدم: توزيع الحالات في مستويات المتغير التابع متجانس عبر مستويات المتغير المستقل

الفرضية البديلة: توزيع الحالات في مستويات المتغير التابع غير متجانس عبر مستويات المتغير المستقل

أو بصيغة رياضية أخرى

$$H_0 : o_{ij} = \varepsilon_{ij}$$

وتمثل $o_{ij}$ التكرارات المشاهدة في حين $\varepsilon_{ij}$ تمثل التكرارات المتوقعة. وتشير فرضية العدم إلى عدم وجود فروق جوهرية بين التكرارات المشاهدة والتكرارات المتوقعة في كل خلية من خلايا الجدول التقاطعي

الفرضية البديلة

$H_a : o_{ij} \neq \varepsilon_{ij}$ (لخلية واحدة على الأقل)

وتشير لفرضية البديلة إلى وجود خلية واحدة على الأقل يختلف فيها التكرارات المشاهدة عن المتوقعة.

مثال:

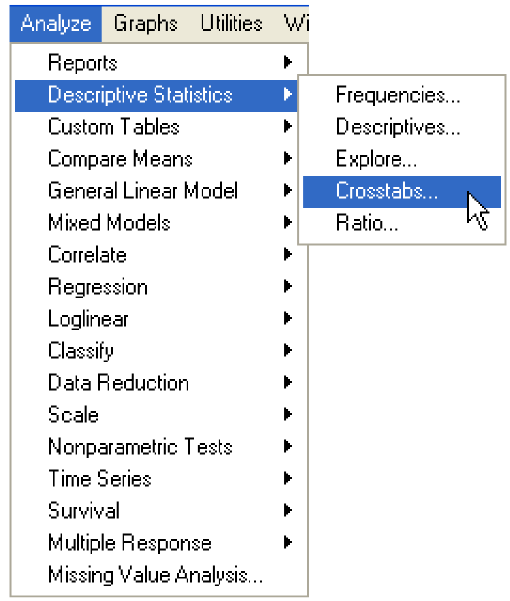
اختبر الفرضية القائلة بأن الجنس والتنصيف الوظيفي متغيران غير مستقلان.

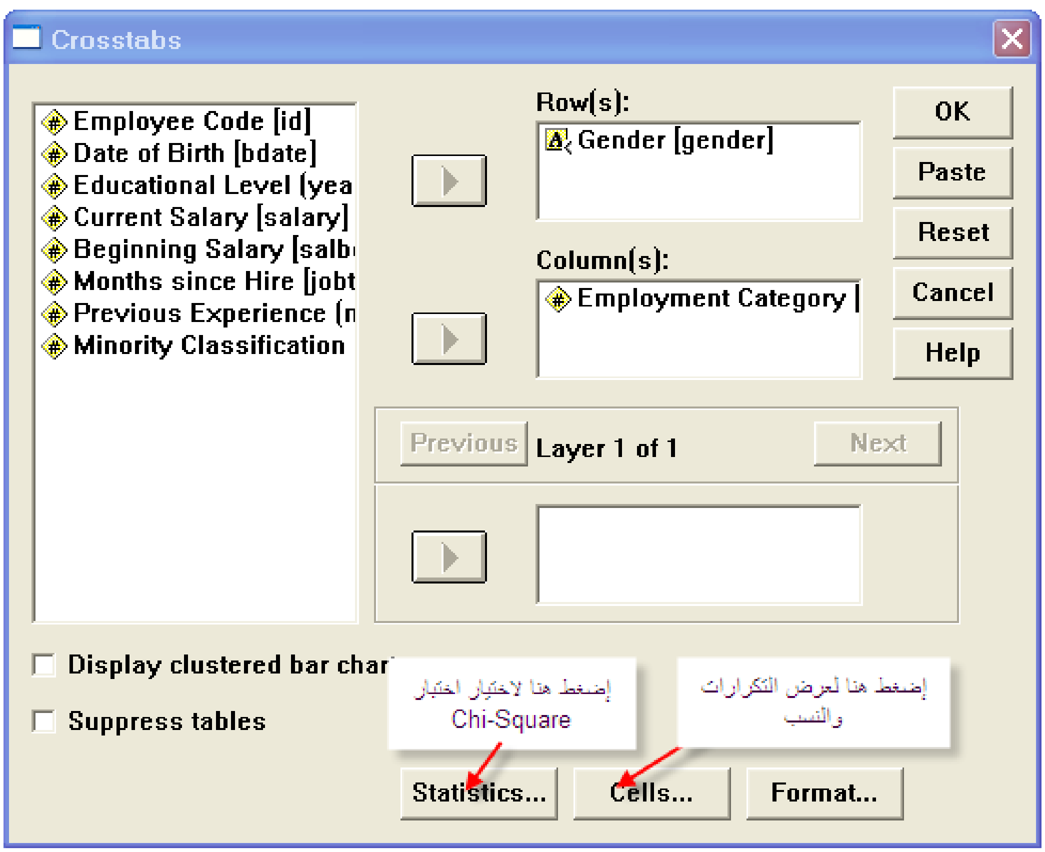
الحل

$H_0$ : الجنس والتصنيف الوظيفي متغيران مستقلان

$H_a$ : الجنس والتصنيف الوظيفي متغيران غير مستقلان.

استخدام برنامج SPSS:



النتائج

**Gender * Employment Category Crosstabulation**

| | | | Employment Category | | | Total |
|---|---|---|---|---|---|---|
| | | | Clerical | Custodial | Manager | |
| Gender | Female | Count | 206 | 0 | 10 | 216 |
| | | Expected Count | 165.4 | 12.3 | 38.3 | 216.0 |
| | Male | Count | 157 | 27 | 74 | 258 |
| | | Expected Count | 197.6 | 14.7 | 45.7 | 258.0 |
| Total | | Count | 363 | 27 | 84 | 474 |
| | | Expected Count | 363.0 | 27.0 | 84.0 | 474.0 |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 79.277[a] | 2 | .000 |
| Likelihood Ratio | 95.463 | 2 | .000 |
| N of Valid Cases | 474 | | |

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 12.30.

وحيث أن قيمة Asymp. Sig. (2-sided) أقل من $(\alpha = 0.05)$، لذا نرفض فرضية العدم لصالح الفرضية البديلة والتي تشير بأن المتغيران غير مستقلان. وهذا يعني أن هناك تأثير تبادلي بين المتغيرين محل الدراسة.

وبطريقة مشابهة يمكن إجراء اختبار التجانس.