مجلة كلية الفقه

تحليل البيانات المصنفة باستخدام احصاءة (kappa) لجداول التوافق 2*2 مع التطبيق العملي

أ.م.د. مهدي وهاب نعمة نصر الله
كلية الادارة والاقتصاد / جامعة كربلاء
Email:Mehdi.wahab@uokerbala.edu.iq

م.م. محمد طالب خنجر
كلية علوم الحاسوب والرياضيات / جامعة الكوفة
Email:mohammedt.azooz@uokufa.edu.iq

الملخص:

تناولنا في هذا البحث احد الاساليب الاحصائية في تحليل البيانات المصنفة (Categorical Data) عندما تكون الاستجابة فيها عبارة عن ازواج مرتبه في جداول توافقية وكذلك تم التطرق الى القوانين والصيغ الرياضية الخاصة بهذه الاحصائية . حيث تم تطبيق هذا المقياس عملياً على بيانات لاحد المعاهد في العراق من خلال دراسة معدلات تخرجهم وعلاقتها بمعدلات القبول الخاصة بهم أوتم تحليلها احصائياً باستخدام برنامج الاكسل لسهولة تعامله مع المعادلات الرياضية وكانت النتائج تشيرالى وجود فروق معنوية بين الظاهرتين قبل وبعد التخرج استناداً الى معيار الاختبار (kappa Statistic) .

الكلمات المفتاحية:
البيانات المصنفة ، جداول التوافق .

1-المقدمة Introduction

لقد تناولت الكثير من الدراسات السابقة مفهوم البيانات المصنفة (Categorical Data) لما لها من اهمية كبيرة في معرفة درجة التوافق (Association) بين البيانات التي تكون فيها الاستجابة عبارة عن ازواج مرتبطة (غير مستقلة)،حيث

حصل تطورٌ هائلٌ في هذا المجال في السنوات السابقة والحالية حتى اصبحت من التطبيقات المهمة في دروس الاحصاء والاحصاء الحياتي . وكذلك نلاحظ ان البيانات المصنفة تهتم بشكل كبير بدراسة المتغيرات التي يمكن تصنيف استجابتها بشكل صفات كما في دراسة استطلاع الرأي (نعم/ كلا) الدراسات البايولوجية بعد أعطاء معالجة (Neg./. Pos)وايضا في التخصصات الطبيه حول نسبة الاصابة (قليله / متوسطه / كبيره)وغيرها من الدراسات الاخرى، حيث أن مشاهده الأستجابة في الوحده التجريبية تسجل بشكل عدد ولذلك تسمى في بعض المصادر العلمية بالبيانات المعدوده وان هذه البيانات تمتلك توزيعات أحصائية مختلفة (Binomial dist.,. Poisson dist., .Multinomial dist.)

تم تطبيق هذا المقياس على بيانات حقيقية مأخوذة من احد المعاهد العراقية والتي تمثل معدلات تخرجهم من المعهد مقارنةً مع معدلات تخرجهم من الاعدادية

2- هدف البحث

يهدف البحث الى تسليط الضوء على كيفية استخدام هذا المقياس احصائية (kappa Statistic) في تحليل البيانات المصنفة بين متغيرين والموضوعة في جدول توافقي (2*2) وكيفية تفسير النتائج النهائية التي من خلالها كيف يتم معرفة مدى قوة التوافق او العلاقة بين المتغيرات الموضوعة قيد الدراسة .

3-الجانب النظري[2,3,6,7]

3-1 البيانات المصنفة

بدأ التطور الهائل في تحليل البيانات المصنفة خلال عقد الستينات المنصرم واستمرت الدراسات عنه بشكل كبير جداً في السنوات اللاحقة والفترة الحالية وذلك بسبب الحركة السريعة في التطور الذي حصلت في العالم والذي يعد في هذا الوقت قرية صغيره نتيجه الخدمات والاتصالات المتوفره والواسعة الانتشار إذ شمل هذا التطورموضوع البيانات المصنفه والتي اصبحت في الوقت الحالي تطبق بشكل واسع في كثير من دول

العالم المتطورة ويشير هذا المصطلح (البيانات المصنفة) الى البيانات التي يمكن وضعها او تمثيلها بشكل تكرار يقع في مجاميع تسمى اصناف (Categories) وإذ يعرف المتغير المصنف بانه االتصنيف او التقسيم المناسب للمتغير المقاس . والجدول الذي يتضمن هذه التكرارات يسمى بجدول التوافق من درجة معينة كأن يكون (2*2) او (2*3) او غير ذلك . ولتوضيح الفكرة اكثر يمكن تلخيصها كما في الجدول التالي:

جدول رقم (١) يبين مخطط لجدول توافق من درجة (r*c) .

		B							Total
		b_1	b_2	b_3	….	J	….	b_c	
	a_1	n_{11}	n_{12}	n_{13}	….	n_{1j}	….	n_{1c}	$n_{1.}$
	a_2	n_{21}	n_{22}	n_{23}	….	n_{2j}	….	n_{2c}	$n_{2.}$
A	a_3	n_{31}	n_{32}	n_{34}	….	n_{3j}	….	n_{3c}	$n_{3.}$
	……	….	….	….	….	n_{31}	….	….	….
	i	n_{i1}	n_{i2}	n_{i3}		n_{ij}		n_{ic}	$n_{i.}$
	……	….	….	….	….	….	….	….	….
	a_r	n_{r1}	n_{r2}	n_{r3}	….	n_{rj}	….	n_{rc}	$n_{r.}$
Total		$n_{.1}$	$n_{.2}$	$n_{.3}$	….	$n_{.j}$	….	$n_{.c}$	$n_{..}$

حيث ان :

n_{ij} : تمثل التكرارات المشاهدة .

$n_{i.}$: تمثل مجموع التكرارات المشاهدة للصف i .

$n_{..}$: مجموع التكرارات الكلية .

2-3 جداول التوافق من رتبة (2*2)[1,4,5,8]

يهتم موضوع البيانات المصنفة بالجداول المبنية على اساس متغيرين او اكثر معبر عنهما ببيانات على شكل مجموعات او اصناف تسمى بجداول التوافق (Contingency Table) . وهذه المتغيرات تكون اما نوعية او كمية.

ولتوضيح هذا النوع من الجداول يكون كما يلي :

ولتوضيح هذا النوع من الجداول يكون بالشكل الآتي :

جدول رقم (٢) يمثل جدول التوافق من درجة (2*2)

		B		Total
		b_1	b_2	
A	a_1	a	b	a+b
	a_2	c	d	c+d
Total		a+c	b+d	N

إذان المتغير الاول والثاني كلاهما مصنفان الى نوعين ، وان (a,b,c,d) تمثل التكرارات المشاهدة بين المتغيرين .

ولحساب التكرارات المتوقعة لكل خلية في الجدول اعلاه تكون حسب الصيغ الرياضية التالية

التكرار المتوقع للخلية الاولى= $\frac{(a+b)(a+c)}{n}$

التكرار المتوقع للخلية الثانية= $\frac{(a+b)(b+d)}{n}$

التكرار المتوقع للخلية الثالثة= $\frac{(c+d)(a+c)}{n}$

التكرار المتوقع للخلية الرابعة= $\frac{(c+d)(b+d)}{n}$

بما ان الجدول التوافقي من النوع (2*2) لذا فان هناك اربع تكرارات متوقعة ومحسوبة بالصيغ الرياضية اعلاه . وبتطبيق الصيغة الرياضية لكاي سكوير لاختبار فرضية الاستقلالية فاذا كانت قيمة كاي سكوير المحسوبة اكبر من قيمة كاي سكوير الجدولية بمستوى معنوية معينة فان هذا يدل على وجود علاقة بين الاصناف وعلى العكس من ذلك فانه يدل على عدم وجود علاقة بين المتغيرين أي استقلالية المتغيرين .

والصيغة الرياضية لكاي سكوير المحسوبة تكون كالاتي :

$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim X^2_{(\alpha,v)} \; ; v = (r-1)(c-1) \dots\dots\dots\dots\dots\dots (1)$$

حيث ان :

O_{ij}: تمثل قيم التكرارات المشاهدة للخلايا في الصف i والعمود j .

E_{ij} : تمثل قيم التكرارات المتوقعة للخلايا في الصف i والعمود j .

<u>3-3 الصيغ والقوانين الرياضية لمقياس (kappa Statistic)</u> (5,7,9)

يعدُّ مقياس احصائية كابا (kappa Statistic) احدى المقاييس الاحصائية المهمة التي يتم من خلالها التعرف على وجود علاقة او توافق (Association) بين المتغيرين من عدمها . والصيغ الرياضية الخاصة بهذه الاحصائية تكون كما يلي:

جدول رقم (٣) يمثل جدول التوافق من درجة (2*2)

Y				
X		y_1	y_1	المجموع
	x_1	f_{11}	f_{12}	R_1
	x_2	f_{21}	f_{22}	R_2
	المجموع	C_1	C_2	n

حيث ان :

f_{11} , f_{12} , f_{21} , f_{22} : تمثل التكرارات المشاهدة بين **المتغيرين** .

R_1 , R_2 : تمثل مجموع التكرارات المشاهدة للصف الاول والثاني على التوالي .

C_1 , C_2 : تمثل مجموع التكرارات المشاهدة للعمود الاول والثاني على التوالي .

علماً ان هذا المقياس يستخرج من الصيغة الرياضية التالية:

$$\mathrm{K} = \frac{p_0 - p_e}{1 - p_e} \quad \ldots\ldots\ldots \quad (2)$$

علماً بان 0 : يمثل الاحتمال المشاهد للتطابق ويحسب من الصيغة الرياضية التالية:

$$p_0 = \frac{1}{n}(f_{11} + f_{22}) \quad \ldots\ldots\ldots \quad (3)$$

وان p_e : يمثل الاحتمال المتوقع للتطابق ويحسب من الصيغة الرياضية الآتية:

$$p_e = \frac{1}{n^2}(R_1C_1 + R_2C_2) \quad \ldots\ldots\ldots \quad (4)$$

جدول رقم (٤) يبين درجة التوافق استناداً الى مقياس احصاءة كاباء

درجة التوافق بين المتغيرين	ضعيف	جيد	ممتاز
	$K \in [0,0.4]$	$K \in [0.4,0.75]$	$K > 0.75$

اما تباين المقياس (K) فيحسب من القانون التالي:

$$V(K) = \frac{1}{n(1-p_e)^2}(p_e + {p_e}^2 \frac{1}{n^3}\sum_{i=1}^{2}(R_i C_i(R_i + C_i))) \dots\dots\dots\dots\dots\dots (5)$$

والمعيار المستخدم لاختبار الفرضية عند استخدام هذا المقياس هي **احصائية** Z التي تتبع التوزيع الطبيعي بمتوسط صفر وتباين واحد اي ان : $Z \sim N(0,1)$

والفرضية المراد اختبارها هي:

$$H_o : K = 0$$

$$H_1 : K \neq 0$$

ولحساب **احصائية** Z تحت فرضية العدم تكون حسب الصيغة الرياضية التالية :

$$Z = \frac{K}{S_e(K)} \sim Z \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (6)$$

حيث ان $S_e(k)$ الذي يمثل الخطأ المعياري للمقياس K يمكن ايجاده ايضا من خلال الصيغة الرياضية التالية :

$$S_e(K) = \sqrt{\frac{p_0(1-p_0)}{n(1-p_e)^2}} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (7)$$

ثم تقارن قيمة المحتسبة من الصيغة اعلاه مع القيمة الجدولية عند مستوى معنوي معين كأن يكون 0.10 ، 0.5 او غيرها ، فاذا كانت المحتسبة اكبر من الجدولية نرفض فرضية العدم والقبول بالفرض البديل وهذا مؤشر لوجود توافق (association) بين المتغيرين، اما اذا كان المحسوبة اقل من الجدولية فان القرار هو قبول بفرضية العدم وهذا يعني عدم وجود توافق بين المتغيرين .

4 – الجانب التطبيقي : Application Part

تم تطبيق مقياس احصائية كاباء (Kappa Statistic) على بيانات مأخوذة من احد المعاهد العراقية والمتمثلة بمعدلات تخرجهم من المعهد مقارنة مع معدلات تخرجهم من السادس الاعدادي (معدلات قبولهم في المعهد) حيث اعطيت تصنيف (positive) للطالب الذي حصل على معدل 70 فأكثر عند قبوله في المعهد او عند تخرجه منه . وتصنيف (Negative) للطالب الذي حاز على معدل اقل من 70 سواء في تخرجه من المعهد او تخرجه من الاعدادية .

ويمكن اعطاء فكرة اكثر توضيحاً لما سبق من خلال الدالة التالية :

$$f(mean) = \begin{cases} positive & mean \geq 70 \\ \text{Negative} & \text{mean} < 70 \end{cases}$$

وعلى اساس الدالة اعلاه تم تصنيف البيانات المستحصل عليها وتم وضعها في جدول توافقي (2*2) حيث الصفوف كانت تمثل التصنيف عند التخرج من الاعدادية أوالاعمدة تمثل التصنيف عند التخرج من المعهد وكما مبين في الجدول التالي:

جدول رقم (٥) يبين التكرارات المشاهدة والمتقاطعة بين المتغيرين

		Response after		Total
Response Before		Pos.	Neg.	
	Pos.	136	92	228
	Neg.	69	240	309
Total		205	332	537

1-4 التحليل الاحصائي:

تم تطبيق برنامج الاكسل في التحليل الاحصائي (Statistical analysis) وذلك لمرونته بالتعامل مع المعادلات الرياضية باي شكل من الاشكال . وبعد ادخال البيانات الى برنامج الاكسل واستخدام الايعازات الموجودة في البرنامج تم

التوصل الى النتائج المدرجه في ادناها عتماداً على القوانين التي تم التطرق اليها في الجانب النظري والخاصة بالاحصائية قيد الدراسة وكانت كما يلي :

1) $P_o = 0.7$

2) $p_e = 0.518$

3) $K = 0.378$

4) $V(K) = 0.04$

5) $S_e(K) = 0.043$

6) $Z = 8.8 \sim N(0,1)$

7) $P - value = 1 - \emptyset(Z) < 0.001$

ومن خلال ملاحظة النتائج اعلاه نلاحظ ان قيمة هذه الاحصائية قيد الدراسة تشير الى وجود فروق معنوية بين الظاهرتين مما يدل على وجود علاقة بين متغير الصفوق والاعمدة .

5- الاستنتاجات والتوصيات:

5-1 الاستنتاجات :

• من خلال التحليل الاحصائي للبيانات باستخدام برنامج الاكسل في الجانب التطبيقي للبحث تم التوصل الى ان هناك علاقة بين الظاهرتين والتي تم معرفتها من خلال قيمة K و قيمة (P-Value) .

• سهولة هذه الطريقة في تحليل البيانات لمعرفة مدى التوافق بين المجموعات وسريعة بالوصول الى النتائج .

5-2 التوصيات :

من خلال الدراسة اعلاه يوصي الباحث بما يلي :

• زيادة الاهتمام بدراسة البيانات المصنفة (Categorical Data) لانها مهمه وذات تطبيقات كثيرة في العلوم منها العلوم الجغرافية والطبية وعلوم الحياة وغيرها .

References:
1- Agresti, A. 1996 , An Introduction to categorical Data Analysis. NY:Wiley.
2- Alan Agresti , 2007, An Introduction to Categorical Data Analysis, 2nd Edition. John Wiley & Sons, Inc.
3- Agresti. , A. 2002. " categorical Data Analysis" (2nd Edition New York, NY: John Wiley & Sons.
4- Blasuius, J & Greenacre, M. (Editors) 1998. Visualization of Categorical Data. San Diego: Academic Press.
5-Daniel, A. Yu Xie, 2008,Statistical Methods for Categorical Data, 2nd Edition, Emerald Group, UK.
6- Fisher, R. A. 1992. "On The interpretation of Chi-square from contingency table", and the calculation of P. J. Roy. Statist. Soc.
7- LIoyd, C.J. 1999."Statiistical Analysis of Categorical Data". NEW York: Wiley.
8- Powers, D.A. & Xie, Y 1999. Statiistical Method for Categorical Data Analysis. Academic Press.
9- Simonoff, J. S. 2003. " Analysis Categorical Data". NEW York: Springer.

Analyze the data classified using the kappa statistic for 2 * 2 Association tables with the application

Abstract

In this research, we dealt with one of the statistical methods in analyzing the classification Data (Categorical Data) when the response in it is non-independent pairs, as well as the mathematical laws and formulas for this statistics were discussed. Where this measure was applied in practice to real data for one of the institutes in Iraq by studying their graduation rates and their relationship to their admission rates, and they were statistically analyzed using the Excel program for easy handling of mathematical equations and the results indicated that there were significant differences between the two phenomena before and after graduation based on the test criterion (p - Value.(

مجلة
كلية الفقه