جامعة الجزائر 3

كلية الاعلام

السنة الثانية

(المجموعتين 1 و 2)

مقياس: برمجيات تحليل البيانات

المحاضرة الثانية

الأستاذة بن ناصف إيمان

اساليب التحليل الإحصائي للبيانات

1 – التحليل أحادي المتغيرة Analyse univarie

يعد التحليل أحادي المتغيرة من أسهل أنواع الأساليب الإحصائية، حيث يهتم بمتغير واحد فقط، حتى و إن كان عدد المتغيرات كبيرا، فإن هذا النوع من التحليل يدرس كل متغيرة على حدى لمراقبة جودة المعلومات من حيث خلوها من الأخطاء و التعرف على القيم الشاذة في البيانات و تلخيص المعلومات في شكل أرقام أو رسوم بيانية. كما يسمح لنا هذا النوع من التحليل من مقارنة النتائج المحصل عليها بين المتغيرات و تختلف أنواع التحليل باختلاف مقاييس المتغيرات من حيث كونها بيانات إسمية أو ترتيبية أو كمية.

1-1 - مقاييس النزعة المركزية

1-1-1 – المتوسط الحسابي La moyenne arithmétique

و يمثل مركز الثقل في أي البيانات الجغرافية، و الغرض من استخدامه هو الاستغناء عن استقراء مفردات المجموعة كلها، و يحسب بإيجاد مجموع قيم المفردات أو الحالات و قسمته على عددها (عدد المفردات). يعرّف المتوسط الحسابي على أنه مجموع قيم المفردات الخاصة بالمتغير في إطار العينة (x1, x2, x3, ... xn) مقسوم على عدد المفردات n (الحالات) و يرمز له بالرمز X و يحسب بتطبيق المعادلة التالية :

$$\overline{X} = \frac{\sum x_i}{n}$$

x_i = المفردة أو الحالة

n = عدد المفردات

سنأخذ بيانات الجدول التالي كمثال لحساب المتوسط الحسابي بالمتسلسلات الإحصائية البسيطة :

التساقطات المطرية بالميلمتر بعشر محطات مناخية من المغرب

المحطة المناخية	التساقطات (ملم)
الحسيمة	380
الدار البيضاء	445
أكادير	230
الصويرة	222
فاس	523
مراكش	172
الرباط	560
وجدة	226
الناظور	443
الجديدة	409
n= 10	$\sum x_i$ = 3610

تتكون هذه المتسلسلة الإحصائية من 10 حالات cases، وتمثل "التساقطات السنوية" المتغيرة المدروسة بهذه العينة، و من الملاحظ أن التساقطات تتراوح ما بين 172 ملم كأدنى معدل و 560 كمعدل أعلى.

إذا أردنا حساب متوسط التساقطات بالمحطات المناخية العشر فسنقوم بقسمة مجموع كمية التساقطات بالمحطات على عشرة

أي

$$\overline{X} = \frac{\sum x_i}{n} = \frac{3610}{10} = 361$$

إذن متوسط التساقطات بالمجال المدروس هو 361 ملم

أما في حالة المتسلسلات الإحصائية المبوبة (بيانات عبارة عن فئات) حيث كل فئة يقابلها عدد من التكرارات فإن المتوسط الحسابي يحسب كالتالي:

$$\overline{X} = \frac{\sum F_i x_i}{\sum F_i}$$

x_i = مركز الفئة

F_i = تكرار الفئة

$\sum F_i$ أو n = التكرار الكلي

سنأخذ بيانات الجدول التالي كمثال لتوضيح كيفية التعامل مع الفئات و كيفية حساب مركز الفئة

توزيع مساحة الاستغلاليات الزراعية بالهكتار بالنسبة ل 281 استغلالية

حجم الاستغلالية بالهكتار	عدد الاستغلاليات الزراعية	x_i	$F_i x_i$
أقل من 4	14	2	28
[4-8[	61	6	366
[8-12[	94	10	940
[12-18[	90	15	1350
أكثر من 18	22	21	462
	$\sum F_i = 281$		$\sum F_i x_i = 3146$

- **حساب مركز الفئة x_i**

مركز الفئة هو المتوسط الحسابي لطرفي الفئة. فمثلا مركز الفئة [4-8[يساوي مجموع الطرفين مقسوم على 2

$$x_{i[4-8[} = \frac{4+8}{2} = 6$$

2-1-1 – الوسيط La médiane

هي القيمة التي تتوسط توزيع القيم في البيانات، حيث تقسم العينة إلى جزأين متساويين، و ذلك بعد ترتيب عناصر المجتمع إما تنازليا او تصاعديا، و يرمز له بالرمز Me و تختلف طرق حساب الوسيط حسب عدد القيم و نوعها.

- في الحالة التي يكون فيها مجموع المفردات عددا فرديا فإن حساب الوسيط يتم بإضافة رقم واحد لمجموع المفردات و قسمة المجموع الكلي على 2

$$\frac{n+1}{2}$$

حيث أن n تمثل عدد أفراد العينة .

سنأخذ قيم الجدول 1 كمثال و سنحذف منه قيمة محطة مراكش و ستصبح المتسلسلة الغحصائية على الشكل التالي بعد ترتيبها تصاعديا

222; 226; 230; 380; 409; 443; 445; 523; 560

بحيث n=9 و بالتالي يصبح الوسيط

$$\frac{n+1}{2} = \frac{9+1}{2} = \frac{10}{2} = 5$$

أي أن قيمة الوسيط هي قيمة n_5 و التي تساوي 409

- في الحالة التي يكون فيها مجموع عدد أفراد العينة عددا زوجيا فإن قيمة الوسيط هي المتوسط الحسابي للمفردتين اللتين ترتيبهما $\frac{n}{2}$ و $(\frac{n}{2}+1)$

أي : $$Me = \frac{\frac{n}{2}+(\frac{n}{2}+1)}{2}$$

نأخذ كمثال قيم المتسلسلة الإحصائية بالجدول الأول و نرتبها تصاعديا فتصبح كالتالي :

172; 222; 226; 230; 380; 409; 443; 445; 523; 560

سنجد بأن n=10 أي أن n عدد زوجي و بالتالي فإن قيمة الوسيط تساوي

$$Me = \frac{\frac{10}{2} + (\frac{10}{2}+1)}{2} = \frac{380+409}{2} = 394{,}5$$

– لكن في المتسلسلات المبوبة (الممثلة بالفئات) فإن الوسيط يتم حسابه من خلال المعادلة التالية :

$$Me = L_0 + \frac{\sum \frac{F_i}{2} - F}{fm} \times I$$

Lo : تمثل الحد الأدنى للفئة الوسيطة

F : التكرار المتجمع الصاعد للفئة قبل الفئة الوسيطة

fm : تكرار الفئة الوسيطة

I : مدى الفئة الوسيطة

نأخذ الجدول التالي كمثال و نحسب فيه قيمة الوسيط.

حجم الاستغلالية بالهكتار	عدد الاستغلاليات الزراعية	مركز الفئة x_i	التكرار الصاعد	التكرار النسبي المتجمع الصاعد	التكرار النسبي المتجمع النازل
أقل من 4	14	2	14	4,98	**100**
[4-8[	61	6	75	26,69	95,02
[8-12[	94	10	169	60,14	73,31
[12-18[	90	15	259	92,17	39,86
أكثر من 18	22	21	281	**100**	7,83
	$\sum F_i = 281$				

لدينا $n = Fi = 281$

إذن

$$\frac{n+1}{2} = \frac{281+1}{2} = 141$$

و من خلال التكرار الصاعد لدينا 169 هي القيمة الأقرب ل 141 و بالتالي فإن الفئة]8-12[هي الفئة الوسيطة، و عليه نحسب قيمة الوسيط بتطبيق المعادلة المناسبة لذلك

أي

$$Me = L_0 + \frac{\sum \frac{F_i+1}{2} - F}{fm} * i = 8 + \frac{141-75}{94} * (12-8) = 10,81$$

إذن قيمة الوسيط تساوي 10,81 و هو أقل نسبيا من الوسيط (11,2) و يمكن حساب هذه القيمة من خلال الرسم البياني بتمثيل التكرارات النسبية المتجمعة الصاعدة و التكرارات النسبية المتجمعة النازلة و تأخذ قيمة الوسيط من خلال إحداثيات نقطة تقاطع منحنيي التكرار.

3-1-1 – المنوال Le Mode

هو القيمة التي لها أكبر تكرار في المتسلسلة الإحصائية و قد يكون للمتسلسلة الإحصائية أكثر من منوال. أما في البيانات المبوبة فإن قيمته تحسب بالمعادلة التالية :

$$M = L_0 + \frac{D_1}{D_1 + D_2} * i$$

Lo : تمثل الحد الأدنى للفئة المنوالية

D1 : الفرق بين تكرار الفئة المنوالية و الفئة التي قبلها

D2 : الفرق بين تكرار الفئة المنوالية و الفئة التي بعدها

i : مدى الفئة المنوالية

في المتسلسلات البسيطة يمثل المنوال القيمة الأكثر تواترا بعد ترتيب المتسلسلات ترتيبا تصاعديا أو تنازليا.

2-1 – مقاييس التشتت Paramètres de dispersion

تبحث مقاييس التشتت في كيفية التعرف على مقدار انتشار البيانات أو تبعثرها، فهي على عكس مقاييس النزعة المركزية التي تتمحور حول قيم مركزية بين مجموعة من المتغيرات، فمقاييس النزعة المركزية وحدها لا تكفي لتقديم فكرة دقيقة عن توزيع البيانات بالمجتمع الإحصائي، و لهذا نلجأ إلى دراسة التوزيع باعتماد مقاييس التشتت و الانتشار للتعرف على مدى تشتت مفردات المتسلسلة الإحصائية حول وسطها الحسابي في العينة، و كلما اقتربت قيم المقاييس من الصفر كلما كان التشتت ضعيفا، و العكس صحيح.
أهم هذه المقاييس الخاصة بدراسة التشتت هي التباين و الانحراف المعياري و معامل التشتت و المدى.

1-2-1 – المدى Range

يعتبر من أبسط مقاييس التشتت إذ يعطينا فكرة عن المدى الذي يمكن أن تتشتت به قيم المتسلسلة الإحصائية، و يقترن بأدنى قيمة و أعلى قيمة، و يرمز له بالرمز E و يتم حسابه بطرح أصغر قيمة في العينة من أكبر قيمة في نفس العينة.

$xi - E = xn$

xn أكبر قيمة في العينة

xi أصغر قيمة في العينة

و توافق قيم xn و xi على التوالي في المتسلسلات المبوبة مركز الفئة الأخيرة و مركز الفئة الأولى

2-2-1 – التباين La Variance

يعرف بأنه المتوسط الحسابي لمربعات الانحرافات عن المتوسط الحسابي، و يصطلح عليه في بعض الأحيان "بتباين المجتمع" و يرمز له بالرمز Var(x) أو $(\sigma x)^2$ و يحسب من خلال المعادلة التالية :

في حالة البيانات البسيطة

$$Var(x) = \frac{\sum(x_i - \bar{x})^2}{n}$$

في حالة المتسلسلات المبوبة :

$$var(x) = \frac{\sum Fi(x_i - \bar{x})^2}{\sum F_i}$$

3-2-1 – الانحراف المعياري Ecart-type

يعطي فكرة عن تشتت القيم عن متوسطها الحسابي، و يعادل الجذر التربيعي للتباين و يصطلح عليه كذلك بتباين العينة، أي أن تباين العينة ecart-type يساوي الجذر التربيعي لتباين المجتمع variance

$$\sigma x = \sqrt{Var(x)} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$$

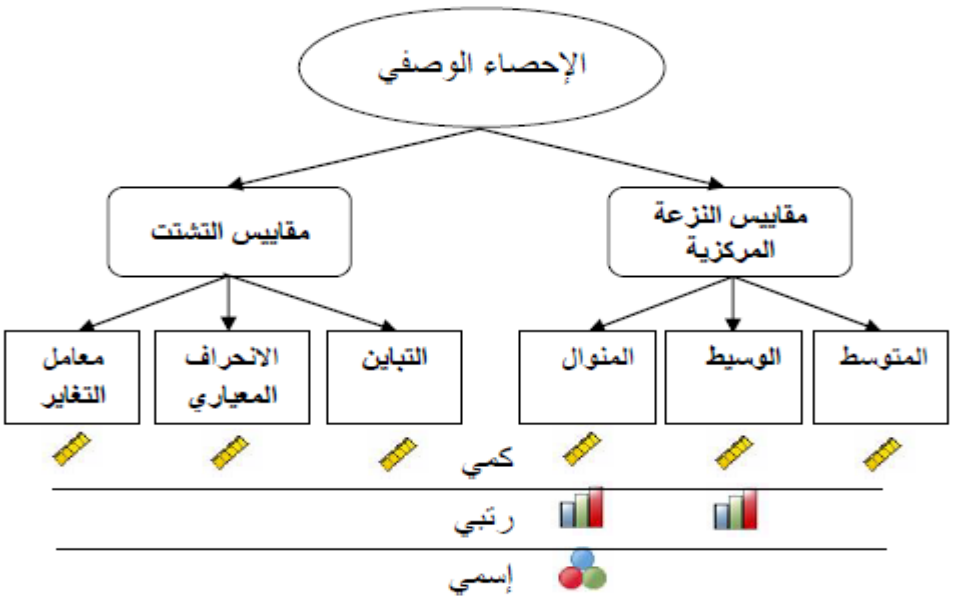
4-2-1 – معامل التشتت Coefficient de variation

يسمى كذلك بمعامل التغير، و يستخدم لتوضيح نسبة تشتت القيم بالعينة المدروسة، و يتم حسابه من خلال المتوسط الحسابي و الانحراف المعياري، إذ يتم قسمة الانحراف المعياري على المتوسط الحسابي و ضربه في 100 للحصول على النسبة المئوية

$$CV = \frac{\sigma x}{\bar{x}} * 100$$

3-1 - أساليب التحليل بمتغير واحد حسب نوع البيانات

الإلمام الجيد بالدراسة يسهل على الباحث تحديد نوع التحليل الذي سيستخدمه، و في كل مستوى من مستويات التحليل يعد معرفة نوع المتغيرات ضروريا لمعرفة نوع المقاييس و الأساليب الإحصائية التي تناسب كل متغيرة، و خاصة عند استخدام البرامج الإحصائية، إذ أنها تسمح بإنجاز قياسات مغلوطة، كحساب المتوسط الحسابي للمتغيرات النوعية، لأنه لطالما يتم ترميز متغير كالجنس مثلا من أجل إخضاعه لعمليات التحليل فإن البرنامج سينفذ جل العمليات على أساس الأرقام التي تم بها الترميز. و فيما يلي تلخيص لأساليب الإحصاء الوصفي حسب نوع المتغيرة.



2 - أساليب التحليل ثنائي المتغيرة

على عكس المستوى الأول المتمثل في التحليل أحادي المتغيرة، الذي يدرس كل متغيرة على حدى، فإن المستوى الثاني يندرج ضمنه التحليل الثنائي الذي ينبني على دراسة العلاقات الثنائية بين متغيرين فقط، حتى و إن فاق عدد المتغيرات 2 فإنه يهتم بدراسة كل متغيرتين على حدى. اهم أنواع التحليل بهذا المستوى و هو الارتباط. و تختلف أساليب التحليل ثنائي المتغيرة حسب نوع البيانات من حيث كونها إسمية أو ترتيبية أو كمية.

1-2 – الارتباط La corrélation

يهتم الارتباط بدراسة العلاقة بين متغيرتين، بحيث إذا تغير أحدهما مال الآخر إلى التغير.

- في الحالة بالارتباط الطردي (ارتباط الموجب) فإن العلاقة بين المتغيرتين هي علاقة طردية موجبة، كلما زادت قيمة المتغيرة الأولى (المتغيرة المستقلة) زادت بالمقابل قيمة المتغيرة الثانية (المغيرة التابعة).
- أما إذا كان التغير في الاتجاه المعاكس، فيسمى بالارتباط العكسي أو ارتباط سالب، اي أن العلاقة بين المتغيرتين هي علاقة عكسية أو سالبة، حيث كلما زادت قيمة المتغير الأول قلت قيمة المتغير التابع أو العكس، كلما قلت قيمة المتغير الأول زادت قيمة المتغير التابع.

يعد الارتباط نوعا من المقاييس الإحصائية الأكثر شيوعا في الدراسات الجغرافية، و خصوصا الارتباط الخطي البسيط لبيرسون Pearson الذي يدرس العلاقة بين متغيرتين كميتين، فكلما اقتربت قيمته من **1** دل ذلك على وجود علاقة ارتباط قوية، و كلما اقتربت من الصفر دلّ ذلك على ضعف أو انعدام العلاقة بين المتغيرتين، فيما أن إشارة معامل الارتباط تدل على نوع العلاقة طردية أو عكسية (موجبة أو سالبة). و يمكن حساب قوة العلاقة بين متغيرتين كميتين x و y و قياس الارتباط بينهما من خلال المعادلة التالية:

$$r = \frac{cov(x,y)}{\sigma x * \sigma y}$$

r : معامل الارتباط

σx : الانحراف المعياري للمتغيرة x

σy : الانحراف المعياري للمتغيرة y

Cov(x,y) : التباين المشترك covariance ل x و y و يتم حسابه باستخدام إحدى المعادلات التالية :

$$cov(x,y) = \frac{\sum(x_i - \bar{x})(y_j - \bar{y})}{n}$$

أو

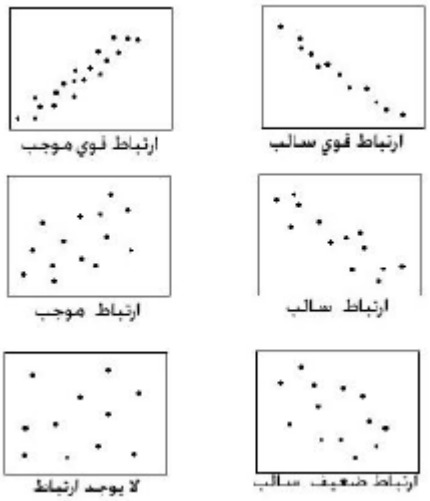
$$cov(x,y) = \left(\frac{\sum(x_i * y_j)}{n}\right) - (\bar{x} * \bar{y})$$

أما في حالة المتسلسلات المبوبة نستخدم المعادلة التالية :

$$cov(x,y) = \frac{\sum F_i(x_i - \bar{x})(y_j - \bar{y})}{\sum F_i}$$

يشترط في الارتباط استقلالية أفراد العينة، و في حالة عدم اعتدال المتغيرتين نستخدم معامل ارتباط آخر (سبيرمان و كندال تاو). ويتم حساب معامل الارتباط البسيط لسبيرمان r من خلال قسمة التباين المشترك ل x و y على حاصل ضرب الانحراف المعياري لكل من x و y

تكون قيمة الارتباط محصورة بين 1+ و 1- و تأخذ العلاقة بين المتغيرتين الأشكال التالية حسب إشارة معامل الارتباط.



2-2 معامل ارتباط الرتب (Spearman)

يعرف بمعامل ارتباط سبيرمان Spearman للرتب، و لذا تختلف قيمته عن قيمة معامل بيرسون (للقيم الأصلية و ليس لرتبها) و هو أقل دقة من معامل ارتباط بيرسون و يتعامل مع البيانات الرقمية و غير الرقمية للترتيب مثل : ضعيف ، متوسط ، جيد .. و يرمز له بالرمز rs و يدخل ضمن الإحصاءات غير المعلمية Non-paramétrique ذات التوزيع الحر و قيمته موجبة أقل أو تساوي الواحد الصحيح، و تحسب قيمته باستخدام المعادلة التالية :

$$r_s = 1 - \frac{\sigma - \sum D^2}{n(n^2 - 1)}$$

n : عدد المشاهدات

D : الفرق بين رتبتي كل قيمتين متقابلتين

و يعتمد معامل الارتباط الرتبي على رتب مستويات المتغيرين للقيم الأصلية من خلال ترتيب مفردات كل متغير من المتغيرات الترتيبية ترتيبا تصاعديا أو تنازليا مع إعطاء كل مفردة قيمة تبين ترتيبها.

3-2 تحليل كاي تربيع X^2

تقوم فكرة مربع كاي على أساس مقارنة البيانات، أي المشاهدات الفعلية التي تمت مشاهدتها من طرف الباحث ببيانات أخرى متوقعة و التي تعبر عن الفرضيات التي وضعها الباحث، فإذا كانت قيمة (x^2) المحسوبة كبيرة، فإن الفرضية الموضوعة غير صحيحة، أما إذا كانت قيمة x^2 صغيرة فإن الفرضية تكون صحيحة لأن الفروق بين التكرارات المشاهدة و المتوقعة تكون قليلة.

و يستخدم كاي تربيع x^2 (مربع كاي) للبيانات المعبر عنها بالتكرار في مستويين أو أكثر كالذكور و الإناث أو مدينة و قرية،.. و هدفه حساب معامل الارتباط لنسبتين أو أكثر لمتغير واحد (أحادي) أو متغيرين تصنيفيين يضم كل منهما مستويين أو أكثر (ثنائي).

و معادلته كاتالي :

$$x^2 = \frac{\sum(Q_i - E_i)^2}{E_i}$$

حيث أن

Qi : التكرار المشاهد الذي نحصل عليه من خلال الجرد الميداني

Ei: التكرار المتوقع = $\frac{\text{مجموع التكرارات الأفقية} \times \text{مجموع التكرارات العمودية}}{\text{المجموع الكلي للتكرارات}}$

n عدد درجات الحرية = (عدد الأسطر -1) (عدد الأعمدة -1)

أما قيمة كاي تربيع الجدولية تستخرج من جدول x^2 حسب درجة الحرية n و مجال الثقة α الذي يحدده الباحث و غالبا ما يكون 0,5 أو 0,1