المحاضرة النظرية الخامسة

تحليل الارتباط Correlation Analysis

إعداد:

د. حيدر الحسن

8/11/2020

- **تعريف:** يعرف تحليل الارتباط Correlation Analysis بأنَّه أحد التحاليل الإحصائية الهامَّة التي تُستخدم بشكل عام لدراسة العلاقة بين المتغيّرات relationship between variables(مثل التحاليل السابقة الذكر كتحليل الانحدار REGRESSION ANALYSIS وتحليل التباين VARIANCE ANALYSIS) **وهو تلازم صفتين تلازما كمّيا.**

- **يستخدم تحليل الارتباط CORRELATION ANALYSIS في الحالات التالية:**
  1. عندما يكون العامل المؤثّر (أي المتغيّر) **من نوع كمّي.**
  2. عندما يكون العامل المؤثّر من نوع **متغيّر احتمالي Random variable.**
  3. **عندما يكون التأثير من جهتين اثنتين** أي أنَّ العامل المؤثر (المتغيّر) يؤثر على الصفة المختارة للظاهرة المدروسة والعكس صحيح، أي يوجد ما يسمى **تأثير متبادل** Interaction ما بين العامل المؤثّر (المتغيّر) والصفة المختارة للظاهرة المدروسة.

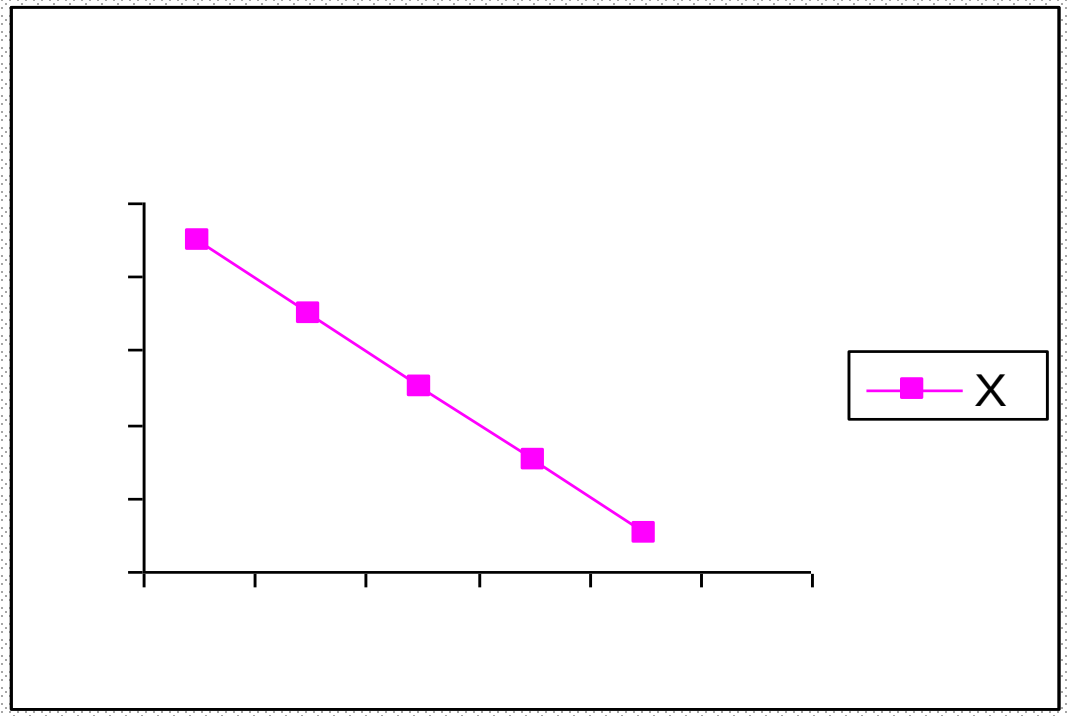
4. **عندما يكون العامل المؤثّر والعامل المتأثّر (التابع) كلاهما صفة مدروسة**، أي عندما ندرس العلاقة بين المتغيّرات الاحتمالية Random variables والتي هي في أغلب الأحيان صفات مدروسة.
5. عندما نرغب في معرفة **قوة العلاقة ما بين العامل المؤثّر والعامل المتأثّر (الصفات المدروسة)**، وذلك من خلال حساب عامل الارتباط كما سنرى بعد قليل.
6. عندما نرغب في معرفة **نسبة التأثير المتبادل INTERACTION ما بين العامل المؤثّر (المتغيّر) والصفة المختارة للظاهرة المدروسة** وذلك من خلال حساب عامل التحديد كما سنرى بعد قليل.
7. إذن إنَّ تحليل الارتباط هو **تلازم صفتين تلازماً أي ارتباطاً كمّياً بين عاملين احتماليين أو صفتين** بحيث إذا طرأ أي تغيّر بالزيادة أو النقصان على إحدى الصفتين يؤدّي بالمقابل إلى تغير بالزيادة أو النقصان للصفة الأخرى.

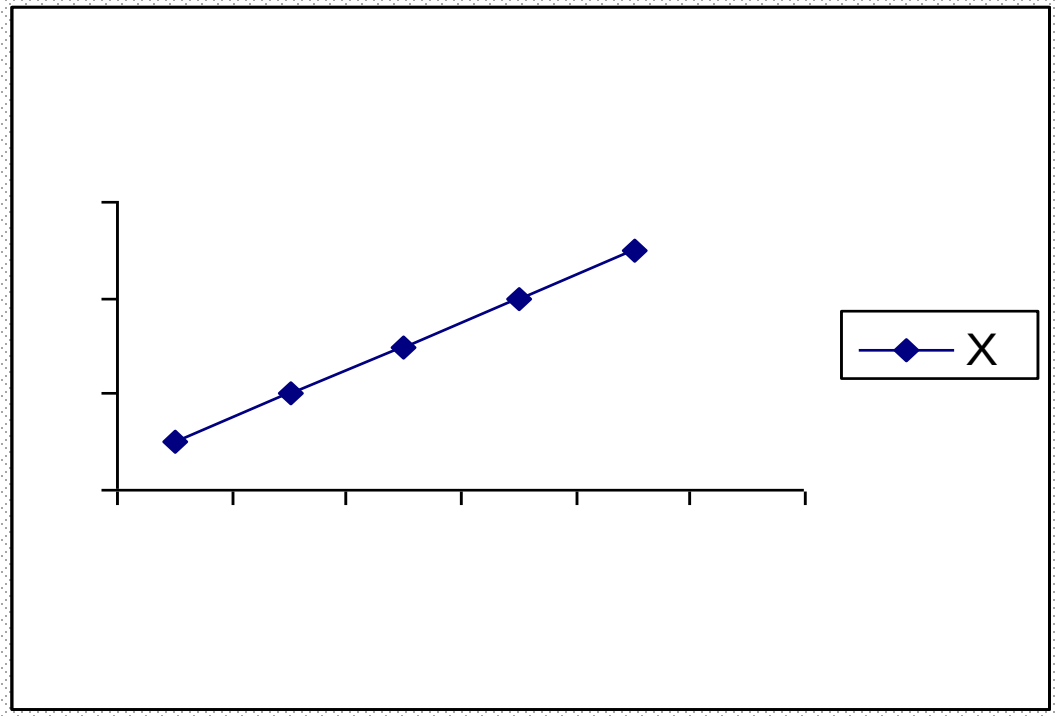
- **بناءً على ذلك يمكن أن نستنتج ما يلي:**
- إذا أدت الزيادة في إحدى الصفتين إلى **زيادة في الصفة الأخرى** نقول إن الارتباط **إيجابي.**
- إذا أدت الزيادة في إحدى الصفتين إلى **نقص في الصفة الأخرى** نقول إن الارتباط **سلبي.**

- إنَّ نتيجة تحليل الارتباط CORRELATION ANALYSIS هي قيمة وحيدة مهمة ألا **وهي عامل الارتباط ويرمز له بـ (r)** وهنا نميز الحالات التالية :
- إذا كانت قيمة عامل الارتباط (r) = 1 نقول أنَّ الارتباط كامل وإيجابي.
- إذا كانت قيمة عامل الارتباط (r) = 1- نقول أنَّ الارتباط كامل وسلبي
- قد يكون معدوم

- إنَّ قيمة **عامل الارتباط 1-< (r) < 1** وبنفس الوقت قيمة **عامل الارتباط (r) <> 0** **نقول أنَّ الارتباط حسب قيمته كما يلي:**
- إذا كانت قيمة عامل الارتباط (r) **بين القيمتين 0 و0.25** **نقول أنَّ الارتباط ضعيف** (موجب أو سالب حسب إشارته).
- إذا كانت قيمة عامل الارتباط (r) **بين القيمتين 0.25 و0.50** **نقول أنَّ الارتباط متوسط** (موجب أو سالب حسب إشارته).
- إذا كانت قيمة عامل الارتباط (r) **بين القيمتين 0.5 و0.75** **نقول أنَّ الارتباط قوي** (موجب أو سالب حسب إشارته).
- إذا كانت قيمة عامل الارتباط (r) **بين القيمتين 0.75 و 1** نقول أنَّ الارتباط **قوي جداً** (موجب أو سالب حسب إشارته).

والأشكال التالية تبيّن الحالات السابقة الذكر بشكل بياني:





أنواع الارتباط:

1- الارتباط البسيط(بيرسون) Simple Correlation Analysis :

هو أبسط أشكال الارتباط المستخدمة في التحليل الإحصائي للبيانات الإحصائية DATA. في هذا الشكل المبسَّط يتم حساب عامل الارتباط (r) باستخدام **إما العلاقة التالية:**

$$r = \sum_{1}^{n} (x_i - \bar{x}) * (y_i - \bar{y}) / \sqrt{\sum_{i}^{n} (x_i - \bar{x})^2 * \sum_{i}^{n} (y_i - \bar{y})^2}$$

أو باستخدام العلاقة التالية التي تسمى عادة طريقة تربيع القيم:

$$r = (\sum (x_i * y_i) - (\sum x_i * \sum y_i)/n) / \sqrt{\left[\sum x_i^2 - (\sum x_i)^2/n\right] * \left[\sum y_i^2 - (\sum y_i)^2/n\right]}$$

حيث أنَّ: r عامل الارتباط

n عدد عناصر العينة

y_i القيم الحقيقية أو التجريبية للعامل الأول أو الصفة الاولى

x_i القيم الحقيقية أو التجريبية للعامل الثاني أو الصفة الثانية

$\bar{x}$ المتوسط الحسابي للعامل الأول أو الصفة الاولى

$\bar{y}$ المتوسط الحسابي للعامل الثاني أو الصفة الثانية

- مثــال: عند دراسة العلاقة ما بين صفة عدد الحبوب في السنبلة الواحدة وصفة الإنتاج لأحد المحاصيل وهو القمح عالي الإنتاج فكان لدينا ما يلي:

الإنتاج	عدد الحبوب/السنبلة
Y	X
5	9
13	7
16	5
23	3
33	1

- ثم أجري التحليل الإحصائي للبيانات الإحصائية DATA فحصلنا على النتائج التالية:

تحليل الارتباط:المؤشرات	
Correlation Statistics	
عامل الارتباط r	0.986063603
R Square	0.972321429
Adjusted R Square	0.963095238
Standard Error	2.033060091
Observations	5

وكما يبدو من الجدول المذكور أعلاه نجد أن:

عامل الارتباط r = 0.986063603

وهنا يمكن أن نستنتج أن الارتباط ايجابي قوي جدا حسب الاعتبارات السابقة الذكر.

2- ارتباط الرتب:(سبيرمان) يستخدم ارتباط الرتب لدراسة العلاقة بين المتغيّرات الاحتمالية Random variables عندما تكون بيانات هذه العوامل مرتبة أو يمكن ترتيبها في رتب معينة. **في هذا الشكل يتم حساب عامل الارتباط (r) باستخدام العلاقة التالية:**

$$r = 1 - (6 * \sum (x_i - y_i)^2) / n(n^2 - 1)$$

- حيث أنَّ:
- y_i رتب العامل الأول أو الصفة الأولى
- x_i رتب العامل الثاني أو الصفة الثانية
- **مثـــال: لدينا ستة طلاب لديهم الدرجات التالية في الإحصاء والاقتصاد:**

الطالب	الإحصاء	الاقتصاد
1	52	46
2	34	57
3	64	53
4	57	26
5	57	43
6	36	56

الحـــل:

نقوم بفرز البيانات تصاعديا فنحصل على الجدول التالي:

الاقتصاد	رتب الاقتصاد	الإحصاء	رتب الإحصاء
26	6	34	6
43	5	36	5
46	4	52	4
53	3	57	2.5
56	2	57	2.5
57	1	64	1

بالنسبة لمادة الإحصاء نلاحظ أنَّها لها أعلى قيمة لذلك أعطيت المرتبة الأولى أمَّا الدرجة 57 فهي مكررة مرتين لذلك تقع رتبتها في الوسط بين الرتبتين أي (2+3)/2 = 2.5 وهكذا تكتمل الرتب بالنسبة لبقية الدرجات.

الآن نعيد كتابة البيانات كما يلي:

الطالب	رتب الإحصاء	الإحصاء	رتب الاقتصاد	الاقتصاد
1	6	52	6	46
2	5	34	5	57
3	4	64	4	53
4	2.5	57	3	26
5	2.5	57	2	43
6	1	36	1	56

نطبق العلاقة السابقة الذكر لحساب عامل الارتباط (r) فنحصل على التالي:

$$r = 1 - 6 * \Sigma (xi - y_i)/n (n^2 - 1)$$

$$= 1 - 6*(56.5)/6*(36 - 1)$$

$$= 1 - (56.5/ 35)$$

$$= 1 - 1.61$$

$$= - 0.61$$

- نلاحظ من النتيجة السابقة أن **عامل الارتباط r = -0.61** وهذا يعني أنَّ **الارتباط سلبي ولكن قوي.**
- بتعبير **آخر أنَّ ترتيب الطالب في مادة الإحصاء معاكس لترتيبه في مادة الاقتصاد** هذا يقودنا للقول إنه إذا كانت علامة الطالب في مادة الإحصاء مرتفعة فإنَّ علامته ستكون في مادة الاقتصاد منخفضة والعكس صحيح.

3- الارتباط الجزئي: يُستخدم الارتباط الجزئي لدراسة العلاقة بين المتغيّرات الاحتمالية Random variables عندما يوجد عدة متغيّرات **ونرغب بحساب الارتباط بين متغيّرين أو صفتين محددتين علماً أنَّ الارتباط مع بقية المتغيّرات معلومة**.

- فإذا كان لدينا عدة متغيّرات X1,X2,.....Xn ورغبنا في حساب الارتباط بين متغيّرين أو صفتين محددتين X1 & X2 علماً أنَّ الارتباط مع بقية المتغيّرات X3,.....Xn معلومة عندئذ نسمي الارتباط المرغوب **"الارتباط الجزئي"** ولحساب الارتباط الجزئي نستخدم العلاقة العامة التالية:

$$r_{12,3} = r_{12} - r_{13} * r_{23} / \sqrt{(1-r_{13}^{2}) * (1-r_{23}^{2})}$$

حيث أنَّ: $r_{12,3}$ الارتباط الجزئي بين متغيّرين أو صفتين محددتين X1 & X2

r_{12} عامل الارتباط البسيط بين متغيّرين أو صفتين X1 & X2

r_{13} عامل الارتباط البسيط بين متغيّرين أو صفتين X1 & X3

r_{23} عامل الارتباط البسيط بين متغيّرين أو صفتين X2 & X3

- الجدير بالقول أنَّ **علاقة الارتباط الجزئي علاقة رجعية** أي أنَّه لحساب الارتباط الجزئي لعاملين آخرين معلومين يلزمنا حساب عوامل الارتباط البسيط المطلوبة.
- مثال(): لدى دراسة العلاقة ما بين **صفة عدد الحبوب في السنبلة الواحدة، عدد السنابل/النبات الواحد وصفة الإنتاج لأحد المحاصيل** وهو القمح عالي الإنتاج حيث لدينا ما يلي:

الإنتاج	عدد الحبوب/السنبلة	عدد السنابل/النبات
X3	X1	X2
5	9	3
13	7	2
16	5	6
23	3	5
33	1	4

نطبّق العلاقة السابقة الذكر لحساب عوامل الارتباط الجزئية($r_{12,3}, r_{13,2}, r_{23,1}$) فنحصل على التالي:

	X2	X1	X3
X2	1		
X1	-0.5	1	
X3	0.358568583	0.986063603	1

من هذا الجدول السابق أعلاه نجد أنَّ:

الارتباط الجزئي بين المتغيّرين X1 & X2 $r_{12,3}$ = -0.5

الارتباط الجزئي بين المتغيّرين X1 & X3 $r_{13,2}$=0.986063603

الارتباط الجزئي بين المتغيّرين X3 & X2 $r_{12,3}$ =0.358568583

4- الارتباط المتعدّد:

يستخدم الارتباط المتعدّد لدراسة العلاقة بين المتغيّرات الاحتمالية عندما يوجد عدة متغيّرات ونرغب بحساب الارتباط بين متغيّر أو صفة محددة X1 مع بقية لمتغيّراتXn....X2,X3.

فإذا كان لدينا عدة متغيّرات X1,X2,....Xn ورغبنا في حساب الارتباط بين متغيّر أو صفة محدَّدة X1 مع بقية المتغيّرات X3,....Xn؛ **عندئذ نسمي الارتباط المرغوب "الارتباط المتعدد" ولحساب الارتباط المتعدّد نستخدم العلاقة العامَّة التالية:**

$$r_{1,23} = (\sqrt{r_{12}^{\ 2} + r_{13}^{\ 2}} - 2 * r_{12} * r_{13} * r_{23}) / (1 - r^{2}_{23})$$

مثـــال: نستعين بالمثال السابق وهو دراسة العلاقة ما بين **صفة عدد الحبوب في السنبلة الواحدة**، **عدد السنابل/النبات الواحد**، **وصفة الإنتاج** للقمح عالي الإنتاج حيث لدينا ما يلي:

عدد السنابل/النبات	عدد الحبوب/السنبلة	الإنتاج
X2	X1	X3
3	9	5
2	7	13
6	5	16
5	3	23
4	1	33

نطبّق العلاقة السابقة الذكر لحساب عامل الارتباط المتعدّد ($r_{1,23}$) فنحصل على النتائج الواردة في الجدول التالي:

Correlation Statistics	
Multiple R عامل الارتباط المتعدّد	0.998213
R Square	0.996429
Adjusted R Square	0.992857
Standard Error	0.894427
Observations	5

- **كما يبدو من الجدول السابق أعلاه نجد أن:**

$r_{1,23} = 0.998213$ الارتباط المتعدد بين المتغير X1 والمتغيرين X2,X3

وهذا يدل على **وجود ارتباط متعدد ايجابي قوي جدا.**

5-عامل التحديد: **يعتبر عامل التحديد أحد المؤشّرات الإحصائية الهامة جداً حيث يقدّم لنا النسبة المئوية لتأثير أحد العوامل (متغيّر أو صفة)** أو عدة عوامل على عامل واحد عند وجود ارتباط متعدد. وعامل التحديد هو رياضيا عبارة عن مربع عامل الارتباط أي أن:

$$B = r^2$$

- حيث أن: B عامل التحديد
- r عامل الارتباط

- **مثـــال:** نستعين أيضا بالمثال السابق وهو دراسة العلاقة ما بين صفة عدد الحبوب في السنبلة الواحدة، عدد السنابل /النبات الواحد، وصفة الإنتاج للقمح عالي الإنتاج وبعد تحليل البيانات الإحصائية حصلنا على النتائج التالية:

Correlation Statistics	
Multiple R عامل الارتباط المتعدد	0.998213
R Square عامل التحديد	0.996429
Standard Error	0.894427
Observations	5

كما يبدو من الجدول السابق أعلاه نجد أن عامل التحديد B = **99.64%**

وهذا يعني أنَّ كلاً من عامل عدد الحبوب/السنبلة وعدد السنابل/النبات يؤثران على الإنتاج بنسبة 99.64%

أمَّا النسبة الباقية وهي 0.36% من التأثير فتعود إلى عوامل أخرى غير معروفة.

اختبار معنوية عامل الارتباط:

إنَّ المقصود من اختبار معنوية الارتباط هو التأكّد إحصائيا من أن عامل الارتباط لا يساوي الصفر؛ وبتعبير آخر إنَّ اختبار معنوية الارتباط هو رياضياً:

$$r \neq 0$$

ولتحقيق ذلك فإنَّه يتوجّب استخدام اختبار ت **(t-test)** وهو أحد اختبارات المعنوية المعروفة وذلك حسب العلاقة التالية:

$$t = r * \sqrt{n - k /(1 - r^2)}$$

حيث أنَّ: k عدد العوامل الداخلة في الاختبار

r عامل الارتباط

n عدد المشاهدات في الاختبار

- **مثـــال:** نعود ونستعين أيضا بالمثال السابق وهو دراسة العلاقة ما بين صفة عدد الحبوب في السنبلة الواحدة، عدد السنابل/النبات الواحد، وصفة الإنتاج للقمح عالي الإنتاج وبعد تحليل البيانات الإحصائية حصلنا على النتائج التالية:

Correlation Statistics	
Multiple R عامل الارتباط ا لمتعدد	0.998213
R Square عامل التحديد	0.996429
Standard Error	0.894427
Observations	5

نطبق اختبار ت t-test السابق الذكر فنحصل على النتائج التالية:

	Coefficients عامل الارتباط	Standard Error	t Stat	P-value
r_2	-1.2	0.326599	-3.67423	0.066743
r_1	-3.6	0.163299	-22.0454	0.002051

- من الجدول الأخير الوارد أعلاه نلاحظ أنَّ قيمة المؤشر Parameter الناتجة عن استخدام اختبار ت (t-test) بالنسبة لعامل الارتباط r_2 هي كما يلي :
- T-STAT = -22.0454
- *P-value* = 0.066743
- نستنتج من ذلك أنَّ **الارتباط غير معنوي** وبالتالي يكون ارتباط العامل (X1) مع العامل التابع (Y) معنوي غير مؤكد إحصائياً حيث نتبع القاعدة التالية حين نحتاج للاستنتاج فيما إذا كان الارتباط معنوي أم لا كما يلي:
- إذا كانت القيمة الاحتمالية لـ *P-value* أصغر أو تساوي 0.05 **نقول أنَّ الارتباط معنوي** وبالتالي يكون ارتباط العامل المستقل (X) مع العامل (Y) معنوي مؤكَّد إحصائيا.
- أمَّا إذا كانت القيمة الاحتمالية لـ *P-value* أكبر 0.05 نقول أن **الارتباط غير معنوي** وبالتالي يكون ارتباط العامل المستقل (X) مع العامل (Y) غير معنوي غير مؤكد إحصائيا وناتج عن الصدفة.

وبطريقة أخرى لمعرفة معنوية الارتباط:

إذا كانت القيمة الإحصائية T-STAT، المحسوبة كما تسمّى أحيانا، أكبر من قيمة T-TABLE، الجدولية كما تسمى أحياناً، **عند درجة حرية مناسبة نقول أنَّ الارتباط معنوي** وبالتالي يكون ارتباط العامل المستقل (X) على العامل (Y) معنوي مؤكد إحصائياً.

أمَّا إذا كانت القيمة الإحصائية T-STAT أصغر أو تساوي قيمة T-TABLE نقول **أنَّ الارتباط غير معنوي** وبالتالي يكون ارتباط العامل المستقل (X) مع العامل (Y) معنوي غير مؤكد إحصائيا وناتج عن الصدفة.

ـالسلاسل الزمنية: TIME SERIES

- مقدمة: من المعروف أن العديد من الظواهر يتعلّق تطورها بالزمن، وبالتالي فإنَّ دراسة تطوّر مثل هذه الظواهر ومراقبة تغيّراتها وتسجيل المشاهدات لها يتمّ باستخدام ما يسمى **أسلوب السلاسل الزمنية TIME SERIES.**
- تتألف السلاسل الزمنية عادةً من بيانات على شكل جدول مؤلّف من عمودين أو سطرين أحدهما يمثل قيم الزمن والثاني يمثل قيم الظاهرة المدروسة.
- مثــال: لدينا بيانات السلسلة الزمنية التالية التى تشمل تطور صادرات إحدى المعامل مقدرة بآلاف الليرات السورية:

الصادرات	العام
200	1990
150	1991
350	1992
100	1993
400	1994

من الجدير بالتنويه إلى أنّه يوجد بعض الشروط العامَّة للسلاسل الزمنية TIME SERIES نذكر منها ما يلي:

- **وحدة المكان:** أي يجب أن تجمع البيانات الخاصة بالسلسلة الزمنية **من مكان واحد،** أي بتعبير آخر يجب أن يكون عامل المكان **ثابت لا يتغير بتغير الزمان.**
- أي يجب أن تجمع البيانات الخاصة بالسلسلة الزمنية بشرط أن تكون وحدة قياس الزمن واحدة ومتساوية.

بناء على هذا فإنَّ السلاسل الزمنية يمكن أن تنقسم حسب دورية المشاهدات إلى الأنواع التالية:

- سلاسل زمنية سنوية حيث تكون المعطيات أو البيانات الإحصائية DATA سنوية مثل إنتاج القطن السنوي.
- سلاسل زمنية فصلية حيث تكون المعطيات أو البيانات الإحصائية DATA فصلية مثل إنتاج المعمل الربع سنوي.
- سلاسل زمنية أسبوعية حيث تكون المعطيات أو البيانات الإحصائية DATA شهرية مثل إنتاج المعمل أسبوعي.

- **كما أنَّه يمكن تقسيم السلاسل الزمنية TIME SERIES حسب العلاقة بين المشاهدات إلى الأنواع التالية:**

1. **السلاسل الزمنية المترابطة (غير مستقلة):** في هذا النوع من السلاسل تكون كل قيمة من قيم السلسة الزمنية مرتبطة مع القيمة التي تسبقها. أي بتعبير آخر أنَّ كل رقم من أرقام السلسة الزمنية ناتج عن إضافة أو طرح مقدار معين من الرقم السابق.

كمثال على ذلك عدد المتخرجين من الطلاب كمهندسين في القطر العربي السوري حيث أن عدد المتخرجين هذا العام هو عبارة عن مجموع المتخرجين من السنوات الماضية مع عدد المتخرجين هذا العام وهكذا.

2. **السلاسل الزمنية غير المترابطة (مستقلة):** في هذا النوع من السلاسل تكون كل قيمة من قيم السلسة الزمنية مستقلة تماما عن القيمة التي تسبقها.

أي بتعبير آخر أن كل رقم من أرقام السلسة الزمنية مستقل تماما عن أي رقم من الأرقام السابقة.

على ذلك إنتاج الهكتار الواحد من القمح هذا العام.

- **المتوسط الحسابي للسلسلة الزمنية:** يختلف المتوسط الحسابي للسلسلة الزمنية حسب نوع السلسلة كما يلي:
- **المتوسط الحسابي للسلاسل الزمنية المترابطة (غير مستقلة):** يتم حساب المتوسط الحسابي هنا باستخدام العلاقة التالية:

$$\bar{X} = (X_1/2 + X_2 + ...X_n/2)/(n-1)$$

- مثـــال (): لدينا السلسلة الزمنية التالية:

عدد العمال	العام
300	1990
310	1991
320	1992
350	1993
400	1994

والمطلوب حساب المتوسط الحسابي:

الحـــــل: نطبق العلاقة السابقة الذكر أعلاه فينتج لدينا أن:

X` = (300/2 + 310 +.....+400/2)/4

= 1330/4 = **332.5**

- **المتوسط الحسابي للسلاسل الزمنية غير المترابطة (مستقلة) أيضا يتم حساب المتوسط الحسابي هنا باستخدام العلاقة التالية:**

$$\bar{X} = (\sum X) / N$$

- وهنا نجد ان حساب المتوسط الحسابي لهذا النوع سهل جدا وكأنها عينة من الأرقام وتتم عملية الحساب بالتطبيق المباشر كما شرح سابقا في فصل مقاييس النزعة المركزية.

- **مؤشّرات السلاسل الزمنية:**

يقصد بمؤشرات السلاسل الزمنية تلك العمليات الحسابية التي تطبق على السلاسل الزمنية من أجل بيان خصائصها، وتقسم هذه مؤشرات السلاسل الزمنية إلى الأنواع التالية :

1- مؤشرات أساسية: وهي المؤشرات التالية:

- **النمو المطلق:** وهي تعطى من خلال: رقم المقارنة – رقم الأساس
- وتيرة النمو: وهي تعطى من خلال: رقم المقارنة ÷ رقم الأساس ×100
- زيادة النمو: وهي تعطى من خلال: وتيرة النمو -100

- حيث أنّ رقم الأساس هو أي رقم من السلسلة الزمنية ويفضل أن يكون الرقم الأول.

2- مؤشرات متتالية: وهي المؤشرات التالية حيث تتم مقارنة كل عنصر مع الرقم الذي يسبقه.

- مثـــال: لدينا السلسلة الزمنية التالية:

العام	الصادرات	النمو المطلق	وتيرة النمو	زيادة النمو
1990	200	-	100	-
1991	150	-50	75	-25
1992	350	150	175	75
1993	100	-100	50	-50
1994	400	200	200	100

إلى اللقاء في المحاضرة القادمة