



يوم : 2022/05/14

امتحان الدورة العادية في مقياس تحليل البيانات

التمرين الأول:

أسئلة "صح أو خطأ" مع التبرير.

1. تُستعمل الطريقة المعيارية في تحليل المركبات الرئيسية فقط عندما تكون جميع المتغيرات مقاسة بنفس الوحدة .

2. يمكن أن تكون قيم القطر الرئيسي في مصفوفة الارتباط $r(X_j, X_j)$ مختلفة من متغير إلى آخر .

3. في الحالة العادية، مركز الثقل لسحابة النقاط يكون عند المتوسطات الأصلية للمتغيرات $(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p)$.

4. الانحراف المعياري δ للمتغير بعد التمرکز (standardisation) يساوي الصفر .

5. مجموع القيم الذاتية λ الناتجة عن مصفوفة الارتباط يكون دائماً يساوي عدد المتغيرات P.

6. عند إجراء التمرکز على متغير إحصائي، يصبح متوسطه الحسابي \bar{X} مساوياً للواحد.

7. في تحليل ACP normée تكون إحداثيات مركز الثقل على الشكل $(1, 1, 1, \dots, 1)$.

8. مصفوفة الارتباط R ليست مصفوفة متناظرة.

التمرين الثاني:

تم إجراء التحليل بالمركبات الرئيسية (ACP) لمصفوفة الارتباط من 10 بيانات و 3 متغيرات $(var1, var2, var3)$. القيم والاشعة الذاتية المعيارية (normés) لمصفوفة الارتباط هي:

$$u_1 = \begin{pmatrix} -0.72 \\ -0.30 \\ x \end{pmatrix}, \quad u_2 = \begin{pmatrix} 0.15 \\ -0.80 \\ 0.57 \end{pmatrix}, \quad u_3 = \begin{pmatrix} -0.68 \\ y \\ -0.40 \end{pmatrix}$$

المركبات الرئيسية لبعض الافراد موضحة في الجدول التالي:

الجدول 1: المركبات الرئيسية

	comp ₁	comp ₂	comp ₃
Obs ₁	1.45	-1.05	0.88
Obs ₂	-0.40	0.72	-0.61
Obs ₃	0.95	-0.90	0.39
Obs ₄	-0.33	0.80	-0.72
Obs ₅	-0.21	0.07	-0.36

الأسئلة:

1. ما هو التباين الكلي لسحابة النقاط؟
2. احسب القيمة x .
3. اكمل الجدول التالي مع إعطاء العلاقات المستعملة:

الجدول 2: التباينات المفسرة

القيمة الذاتية Valeur propre (λ_j)	Inertie expliquée التباين المفسر	Inertie expliquée cumulée التباين المفسر التجميعي
		0.2067
2.09		

4. احسب جودة تمثيل الفرد الأول بالنسبة للمحور الرئيسي الثاني، مع الاخذ بعين الاعتبار ان جودة تمثيله (الفرد الأول) بالنسبة للمحور الرئيسي الأول تساوي 0.41.
5. بعد إعطاء الصيغة المناسبة، احسب مساهمة الفرد الأول في المركب الرئيسي الثاني.
6. اعط صيغة كتابة المركب الرئيسي comp1 بالأول بدلالة $var1, var2, var3$.
8. اثبت وجود علاقة ارتباط قوية للمتغير 1 في المركب الرئيسي الأول.

التمرين الثالث:

اليك المعطيات التالية

$$f: R^2 \longrightarrow R^2$$

$$(x, y) \longrightarrow (5x+2y, x+4y)$$

1. استخرج المصفوفة M من البيانات السابقة.
2. ماهي القاعدة المعيارية في هذه الحالة.
3. احسب متعدد الحدود المميز للمصفوفة M .
4. احسب القيم والاشعة الذاتية للمصفوفة M .
5. ماذا يساوي التقطيع القطري للمصفوفة M في هذه الحالة.

التمرين الرابع:

تم اجراء تحليل البيانات لجدول يمثل تفصيل ديموغرافي لتغير عدد السكان في الولايات المتحدة حسب كل ولاية (51 ولاية) و 7 متغيرات التالية:

صافي الهجرة الداخلية → Net Domestic Mig.

انتقال/هجرة الفيدراليين والمدنيين من الخارج → Federal/Civilian move from abroad

صافي الهجرة الدولية → Net Int. Migration

Period Births → عدد المواليد خلال الفترة

Period Deaths → عدد الوفيات خلال الفترة

< 65 Pop. Est. → تقدير عدد السكان أقل من 65 سنة

> 65 Pop. Est. → تقدير عدد السكان أكبر من 65 سنة

اليك الجداول والرسومات البيانية التالية

1. ماهو البرنامج المستخدم لاستخراج هذه النتائج؟
2. اعط الاسم المناسب لكل مرحلة (الأولى، الثانية، الثالثة).
3. اعط العنوان المناسب لكل جدول والمخطط.

I. المرحلة الاولى: تحليل

جدول 1:

Variable	Observations	Obs. with missing data	Obs. without missing data	Minimum	Maximum	Mean	Std. deviation
Net Domestic Mig. Federal/Civilian move from abroad	51	0	51	-13,483	27,349	0,246	6,911
Net Int. Migration	51	0	51	-0,293	-0,001	-0,044	0,056
Period Births	51	0	51	0,272	7,879	2,530	1,891
Period Deaths	51	0	51	10,313	20,406	13,864	1,736
< 65 Pop. Est.	51	0	51	4,645	11,896	8,733	1,389
> 65 Pop. Est.	51	0	51	826,278	941,949	874,883	18,494
> 65 Pop. Est.	51	0	51	58,051	173,722	125,117	18,494

اشرح الجدول باختصار:

II. المرحلة الثانية: تحليل

جدول 2:

Variables	Net Domestic Mig.	Federal/Civilian move from abroad	Net Int. Migration	Period Births	Period Deaths	< 65 Pop. Est.	> 65 Pop. Est.
Net Domestic Mig. Federal/Civilian move from abroad	1	0,020	0,206	-0,060	-0,232	0,095	0,095
Net Int. Migration	0,020	1	-0,133	-0,308	0,422	0,377	0,377
Period Births	0,206	-0,133	1	0,295	-0,412	0,204	0,204
Period Deaths	-0,060	-0,308	0,295	1	-0,506	0,640	0,640
< 65 Pop. Est.	-0,232	0,422	-0,412	-0,506	1	0,779	0,779
> 65 Pop. Est.	0,095	-0,377	0,204	0,640	-0,779	1	1,000
> 65 Pop. Est.	-0,095	0,377	-0,204	-0,640	0,779	1,000	1

اشرح الجدول باختصار:

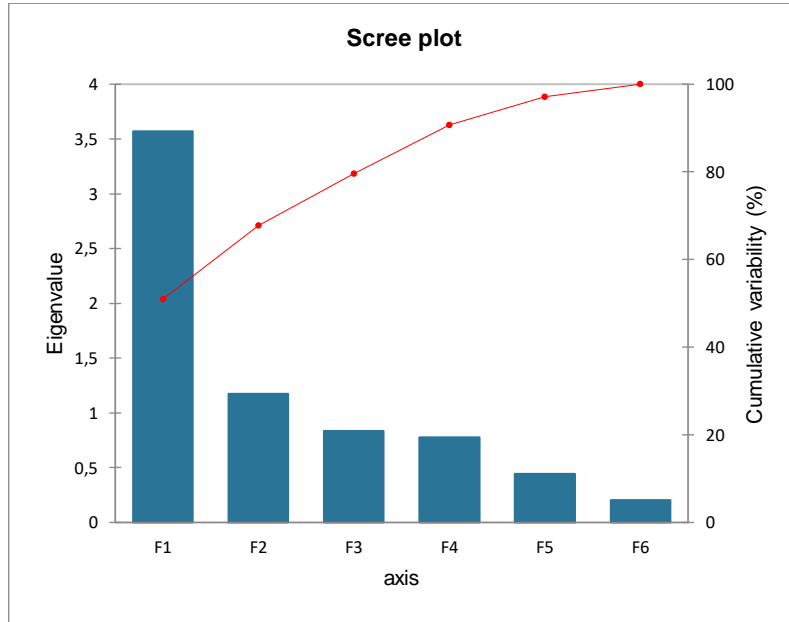
.....III المرحلة الثالثة: تحليل

..... جدول 3:

	F1	F2	F3	F4	F5	F6
Eigenvalue	3,567	1,173	0,835	0,776	0,444	0,204
Variability (%)	50,964	16,756	11,932	11,091	6,342	2,914
Cumulative %	50,964	67,720	79,652	90,744	97,086	100,000

اشرح الجدول باختصار:

.....المخطط 1:



ماذا يمكن ان نستنتج من المخطط

..... جدول 4:

	F1	F2	F3	F4	F5	F6
Net Domestic Mig.	0,085	0,777	-0,458	-0,193	0,373	0,058
Federal/Civilian move from abroad	-0,280	0,195	-0,222	0,896	-0,134	-0,116
Net Int. Migration	0,221	0,520	0,745	0,148	-0,182	0,267
Period Births	0,396	-0,192	0,226	0,309	0,781	-0,222
Period Deaths	-0,468	-0,150	0,047	0,056	0,385	0,778
< 65 Pop. Est.	0,495	-0,122	-0,257	0,140	-0,160	0,359
> 65 Pop. Est.	-0,495	0,122	0,257	-0,140	0,160	-0,359

ماذا تمثل F1, F2, F3 في الجدول 4 :

جدول 5:

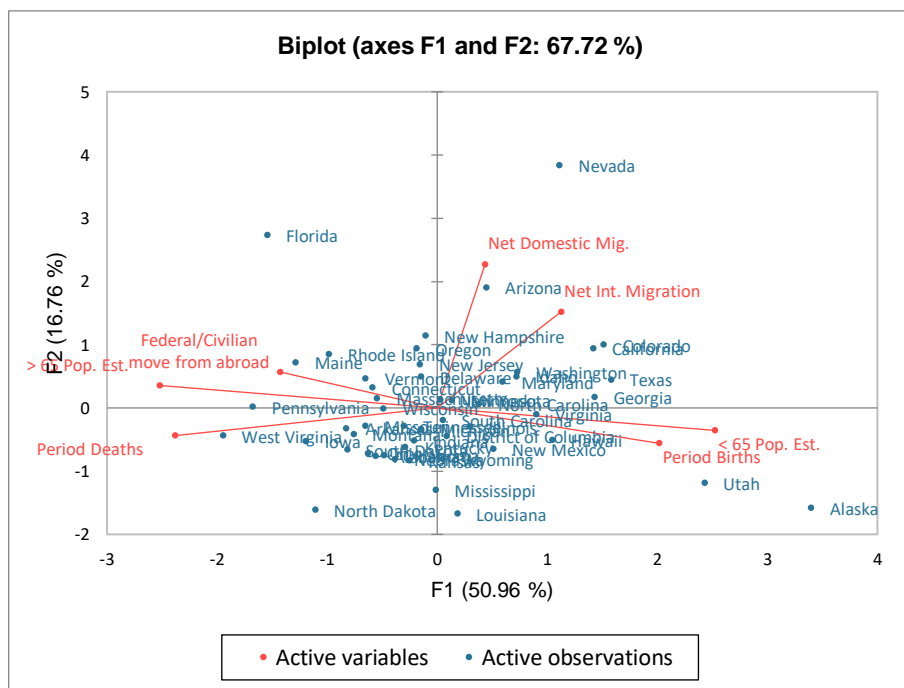
Squared cosines of the variables:

	F1	F2	F3	F4	F5	F6
Net Domestic Mig.	0,026	0,707	0,175	0,029	0,062	0,001
Federal/Civilian move from abroad	0,280	0,044	0,041	0,623	0,008	0,003
Net Int. Migration	0,174	0,317	0,463	0,017	0,015	0,015
Period Births	0,559	0,043	0,043	0,074	0,271	0,010
Period Deaths	0,780	0,026	0,002	0,002	0,066	0,123
< 65 Pop. Est.	0,874	0,017	0,055	0,015	0,011	0,026
> 65 Pop. Est.	0,874	0,017	0,055	0,015	0,011	0,026

Values in bold correspond for each variable to the factor for which the squared cosine is the largest

ماذا تمثل القيم بالخط العريض في الجدول 5 وهل يمكن الاستغناء عن F3, F4 في التمثيل البياني؟ مع التعليل.

المخطط الثنائي لتحليل المكونات الرئيسية (ACP)



ماذا يمثل الخط السابق؟ وماذا يمكن ان تستنتج منه؟

د. مريم بوراس



يوم : 2022/05/14

الحل النموذجي لامتحان الدورة العادية في مقياس تحليل البيانات

التمرين الأول:

أسئلة "صح أو خطأ" مع التبرير.

1. تُستعمل الطريقة المعيارية في تحليل المركبات الرئيسية فقط عندما تكون جميع المتغيرات مقاسة بنفس الوحدة .

خطأ.....

الطريقة المعيارية (ACP normée) تُستعمل غالبًا عندما تكون المتغيرات مقاسة بوحدات مختلفة أو ذات تباينات مختلفة، حتى يتم توحيد المقاييس وإعطاء نفس الأهمية لكل المتغيرات.

2. يمكن أن تكون قيم القطر الرئيسي في مصفوفة الارتباط $r(X_j, X_j)$ مختلفة من متغير إلى آخر .

خطأ.....

في مصفوفة الارتباط تكون عناصر القطر الرئيسي دائمًا: $1=r(X_j, X_j)$

لأن ارتباط المتغير بنفسه يساوي دائمًا 1.

3. في الحالة العادية، مركز الثقل لسحابة النقاط يكون عند المتوسطات الأصلية للمتغيرات $(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p)$.

صح.....

مركز الثقل لسحابة النقاط في الفضاء الإحصائي يساوي متجه المتوسطات الحسابية للمتغيرات.

4. الانحراف المعياري δ للمتغير بعد التمرکز (standardisation) يساوي الصفر .

خطأ.....

بعد التمرکز والتوحيد (standardisation) يصبح: $\delta=1, \bar{X}=0$

أي أن الانحراف المعياري يساوي الواحد وليس الصفر.

5. مجموع القيم الذاتية λ الناتجة عن مصفوفة الارتباط يكون دائمًا يساوي عدد المتغيرات P .

صح.....

لأن أثر مصفوفة الارتباط يساوي عدد المتغيرات.

6. عند إجراء التمرکز على متغير إحصائي، يصبح متوسطه الحسابي \bar{X} مساويًا للواحد.

خطأ.....

بعد التمرکز فقط: $\bar{X}=0$

7. في تحليل ACP normée تكون إحداثيات مركز الثقل على الشكل $(1, 1, 1, \dots, 1)$.

خطأ.....

في ACP normée تكون البيانات متمركزة، وبالتالي: $G=(0,0,\dots,0)$

8. مصفوفة الارتباط R ليست مصفوفة متناظرة.

خطأ.....

مصفوفة الارتباط متناظرة لأن: $r(X_j, X_i) = r(X_i, X_j)$

التمرين الثاني:

1. التباين الكلي في هذه الحالة (التحليل بالمركبات الرئيسية المعياري) يساوي عدد المتغيرات 3.
2. حساب القيمة x:

بما انه في حالة التحليل بالمركبات الرئيسية المعياري الاشعة الذاتية متعامدة اثنين، اثنين، اذن:

$$U_1 * U_2 = 0 \leftrightarrow (-0.72 * 0.15) + (-0.30 * -0.80) + (x * 0.57) = 0$$

$$-0.108 + 0.24 + 0.57x = 0$$

$$x \approx -0.2316$$

3. اكمال الجدول مع إعطاء العلاقات المستعملة:

القيمة الذاتية Valeur propre (λ_i)	Inertie expliquée التباين المفسر	Inertie expliquée cumulée التباين المفسر التجميعي
$\lambda_1/3 = 0.2067 \rightarrow \lambda_1 = 3 * 0.2067 = 0.6201$	0.2067	0.2067
$\lambda_2 = 3 - \lambda_1 - \lambda_3$ $\lambda_2 = 3 - 0.6201 - 2.09$ $\lambda_2 = 0.2899$	$\lambda_2/3 = 0.2899/3 = 0.096633333333$	$0.2067 + \lambda_2/3 = 0.2067 + 0.096633333333 = 0.303333333333$
2.09	$\lambda_3/3 = 2.09/3 = 0.696666666666$	$0.2067 + \lambda_2/3 + 2.09/3 = 0.2067 + 0.303333333333 + 0.696666666666 \approx 1$

4. حساب جودة تمثيل الفرد الأول بالنسبة للمحور الرئيسي الثاني:

$$\cos(\theta_1)^2 = 0.41$$

$$\cos(\theta_2)^2 = ?$$

$$\cos(\theta_1)^2 = (\vec{U}_1 * \overline{ind}_1)^2 / \|\overline{ind}_1\|^2 = 0.41 \rightarrow \|\overline{ind}_1\|^2 = (\vec{U}_1 * \overline{ind}_1)^2 / 0.41$$

$$\cos(\theta_2)^2 = (\vec{U}_2 * \overline{ind}_1)^2 / \|\overline{ind}_1\|^2$$

$$\cos(\theta_2)^2 = (\vec{U}_2 * \overline{ind}_1)^2 * 0.41 / (\vec{U}_1 * \overline{ind}_1)^2 = (-1.05)^2 * 0.41 / (1.45)^2 = 0.21499405469$$

5. حساب مساهمة الفرد الأول في المركب الرئيسي الثاني:

الصيغة:

$$Ctr(ind_1) = \frac{\frac{1}{n}(\overline{ind}_1 * \vec{U}_2)^2}{\lambda_2} = \frac{\frac{1}{10}(-1.05)^2}{0.2899} = 0.38030355294$$

- 6.

a. صيغة كتابة المركب الرئيسي comp1 الأول بدلالة var1, var2, var3

$$Comp1 = -0.72 * var1 - 0.30 * var2 + x * var3$$

$$Comp1 = -0.72 * var1 - 0.30 * var2 - 0.2316 * var3$$

b. علاقة الارتباط بين المتغير 1 والمركب الرئيسي الأول:

$$r(\text{var1}, \text{comp1}) = -0.72 * \sqrt{\lambda_1} = -0.72 * \sqrt{0.6201} = -0.56697428513$$

التمرين الثالث:

اليك المعطيات التالية

$$f: \mathbb{R}^2 \longrightarrow \mathbb{R}^2$$

$$(x, y) \longrightarrow (5x+2y, x+4y)$$

1. استخراج المصفوفة M من البيانات السابقة.

$$M = \begin{pmatrix} 5 & 2 \\ 1 & 4 \end{pmatrix}$$

2. القاعدة المعيارية في هذه الحالة.

$$B_e = \{(1,0) | (0,1)\}$$

3. متعدد الحدود المميز للمصفوفة M.

$$\text{Det}(M - \lambda I) = \begin{vmatrix} 5 - \lambda & 2 \\ 1 & 4 - \lambda \end{vmatrix} = \lambda^2 - 9\lambda + 18$$

4. القيم والأشعة الذاتية للمصفوفة M.

$$\lambda_1 = 6 \rightarrow \vec{U}_1 = (2, 1)$$

$$\lambda_2 = 3 \rightarrow \vec{U}_2 = (-1, 1)$$

5. التقطيع القطري للمصفوفة M في هذه الحالة.

$$D = \begin{pmatrix} 6 & 0 \\ 0 & 3 \end{pmatrix}$$

التمرين الرابع:

تم إجراء تحليل البيانات لجدول يمثل تفصيل ديموغرافي لتغير عدد السكان في الولايات المتحدة حسب كل ولاية (51 ولاية) و 7 متغيرات التالية:

Net Domestic Mig. → صافي الهجرة الداخلية

Federal/Civilian move from abroad → انتقال/هجرة الفيدراليين والمدنيين من الخارج

Net Int. Migration → صافي الهجرة الدولية

Period Births → عدد المواليد خلال الفترة

Period Deaths → عدد الوفيات خلال الفترة

< 65 Pop. Est. → تقدير عدد السكان أقل من 65 سنة

> 65 Pop. Est. → تقدير عدد السكان أكبر من 65 سنة

اليك الجداول والرسومات البيانية التالية

1. البرنامج المستخدم لاستخراج هذه النتائج **XLSTAT**
2. اعط الاسم المناسب لكل مرحلة (الأولى، الثانية، الثالثة).
3. اعط العنوان المناسب لكل جدول والمخطط.

I. المرحلة الاولى: تحليل احادي المتغير

جدول 1: الاحصاءات الوصفية.....

Variable	Observations	Obs. with missing data	Obs. without missing data	Minimum	Maximum	Mean	Std. deviation
Net Domestic Mig. Federal/Civilian move from abroad	51	0	51	-13,483	27,349	0,246	6,911
Net Int. Migration	51	0	51	-0,293	-0,001	-0,044	0,056
Period Births	51	0	51	0,272	7,879	2,530	1,891
Period Deaths	51	0	51	10,313	20,406	13,864	1,736
< 65 Pop. Est.	51	0	51	4,645	11,896	8,733	1,389
> 65 Pop. Est.	51	0	51	826,278	941,949	874,883	18,494
> 65 Pop. Est.	51	0	51	58,051	173,722	125,117	18,494

اشرح الجدول باختصار:

...يمثل الجدول الإحصاءات الوصفية للمتغيرات المتمثلة في (القيمة الصغرى، القيمة الكبرى، المتوسط، والانحراف المعياري) لكل متغير مع تسجيل أكبر متوسط للمتغير < 65 Pop. Est. وايضا أكبر انحراف معياري للمتغيرين < 65 Pop. Est. و > 65 Pop. Est. ما يدل على تشتت عالي في هذين المتغيرين....

II. المرحلة الثانية: تحليل ثنائي المتغير

جدول 2: مصفوفة الارتباط.....

Variables	Net Domestic Mig.	Federal/Civilian move from abroad	Net Int. Migration	Period Births	Period Deaths	< 65 Pop. Est.	> 65 Pop. Est.
Net Domestic Mig. Federal/Civilian move from abroad	1	0,020	0,206	-0,060	-0,232	0,095	0,095
Net Int. Migration	0,020	1	-0,133	-0,308	0,422	0,377	0,377
Period Births	0,206	-0,133	1	0,295	-0,412	0,204	0,204
Period Deaths	-0,060	-0,308	0,295	1	-0,506	0,640	0,640
< 65 Pop. Est.	-0,232	0,422	-0,412	-0,506	1	0,779	0,779
> 65 Pop. Est.	0,095	-0,377	0,204	0,640	-0,779	1	1,000
> 65 Pop. Est.	-0,095	0,377	-0,204	-0,640	0,779	1,000	1

اشرح الجدول باختصار:

..... يمثل الجدول مصفوفة الارتباط (خصائصها: قيم قطرها تساوي 1، متناظرة، قيمها منحصرة بين -1 و 1)، أكبر قيمة 0,779 تدل على ارتباط إيجابي قوي بين المتغيرين، وأصغر قيمة -0,095 تدل على عدم وجود ارتباط بين المتغيرين (ارتباط عكسي ضعيف جدا).....

III. المرحلة الثالثة: تحليل متعدد المتغيرات (التحليل بالمركبات الرئيسية).....

جدول 3: القيم الذاتية.....

	F1	F2	F3	F4	F5	F6
Eigenvalue	3,567	1,173	0,835	0,776	0,444	0,204
Variability (%)	50,964	16,756	11,932	11,091	6,342	2,914
Cumulative %	50,964	67,720	79,652	90,744	97,086	100,000

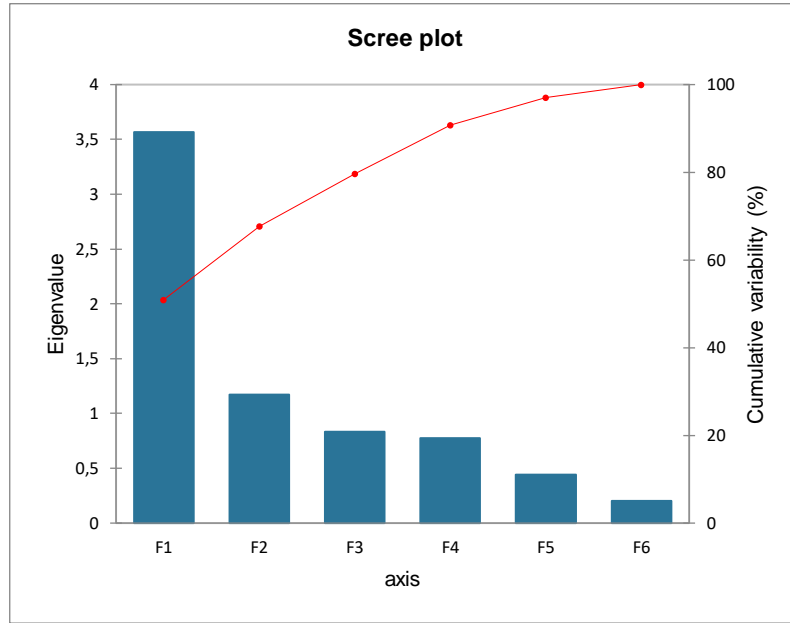
اشرح الجدول باختصار:

..... عدد القيم الذاتية لمصفوفة الارتباط أقل من عدد المتغيرات.....

.....ثاني سطر يتمثل في نسبة التباين $100 * P / \lambda$

.....ثالث قيمة تتمثل في التباين التجميعي

المخطط 1: التمثيل البياني للقيم الذاتية



ماذا يمكن ان نستنتج من المخطط

.....كل عمود يمثل قيمة ذاتية، والخط البياني يمثل التباين التجميعي

كما لدينا F1 و F2 اكبر من 1

جدول 4: الاشعة الذاتية

	F1	F2	F3	F4	F5	F6
Net Domestic Mig.	0,085	0,777	-0,458	-0,193	0,373	0,058
Federal/Civilian move from abroad	-0,280	0,195	-0,222	0,896	-0,134	-0,116
Net Int. Migration	0,221	0,520	0,745	0,148	-0,182	0,267
Period Births	0,396	-0,192	0,226	0,309	0,781	-0,222
Period Deaths	-0,468	-0,150	0,047	0,056	0,385	0,778
< 65 Pop. Est.	0,495	-0,122	-0,257	0,140	-0,160	0,359
> 65 Pop. Est.	-0,495	0,122	0,257	-0,140	0,160	-0,359

ماذا تمثل F1, F2, F3 في الجدول 4 :

...تمثل الاشعة الذاتية المرافقة للقيم الذاتية $\lambda_1, \lambda_2, \lambda_3$

جدول 5:

Squared cosines of the variables:

	F1	F2	F3	F4	F5	F6
Net Domestic Mig.	0,026	0,707	0,175	0,029	0,062	0,001
Federal/Civilian move from abroad	0,280	0,044	0,041	0,623	0,008	0,003
Net Int. Migration	0,174	0,317	0,463	0,017	0,015	0,015
Period Births	0,559	0,043	0,043	0,074	0,271	0,010
Period Deaths	0,780	0,026	0,002	0,002	0,066	0,123
< 65 Pop. Est.	0,874	0,017	0,055	0,015	0,011	0,026
> 65 Pop. Est.	0,874	0,017	0,055	0,015	0,011	0,026

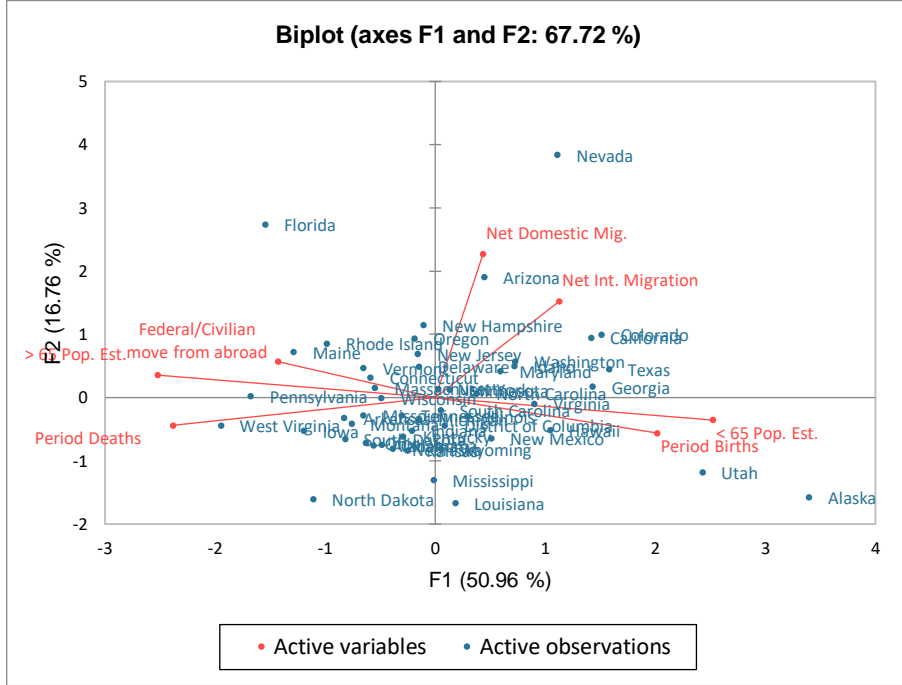
Values in bold correspond for each variable to the factor for which the squared cosine is the largest

ماذا تمثل القيم بالخط العريض في الجدول 5 وهل يمكن الاستغناء عن F3, F4 في التمثيل البياني؟ مع التعليل.

...تمثل القيم بالخط العريض القيم التي لها أعلى جودة تمثيل....

....نعم يمكن الاستغناء عن F3 و F4 لأن نسبة البيانات المحفوظ عليها من F1 و F2 تساوي 67.72% وهي أكبر من 50%....

المخطط الثنائي لتحليل المكونات الرئيسية (ACP)



ماذا يمثل الخط السابق؟ وماذا يمكن ان تستنتج منه؟

....يمثل هذا المخطط الثنائي (Biplot) نتائج تحليل المكونات الرئيسية (ACP)، حيث يوضح العلاقة بين المتغيرات والولايات الأمريكية على المحورين F1 و F2 اللذين يفسران 67.72% من التباين الكلي. نلاحظ وجود ارتباط إيجابي بين متغيرات الهجرة والنمو السكاني، بينما تظهر بعض الولايات مثل Nevada و Alaska بعيدة عن باقي الولايات مما يدل على امتلاكها خصائص ديموغرافية مميزة. كما يبين المخطط أن المحور الأول هو الأكثر أهمية لأنه يفسر أكبر نسبة من المعلومات.....

د. مريم بوراس