Олон хэмжээст өгөгдлийн статистик шинжилгээ хичээлийн нэмэлт жишээ

Хөнгөн атлетикийн гүйлтийн төрлөөрх улс орнуудын үндэсний рекорд амжилт

Г.Махгал

2022 оны 4 сарын 11

Удиртгал

Хөнгөн атлетикийн гүйлтийн төрөлд тамирчид 100, 200, 400, 800, 1500, 5000, 10000 метр гэх мэтчилэн олон төрлийн зайд уралддаг. Хөндлөнгөөс харахад уралдах зайн энэхүү ангилал ихэдсэн мэт санагдана. Иймд спортын энэхүү төрөлд харгалзах өгөгдөл дээр корреляцийн шинжилгээ болон гол хэсгийн шинжилгээ хийж дээр дурдсан уралдах зайн ангиллуудын холбоо хамаарал болон эдгээр олон төрлийн зайн мэдээлэл угтаа хэчнээн хүчин зүйлээр тодорхойлогдох цаашилбал тэдгээр олон төрлийн зайн уралдааныг цөөлж болох эсэх буюу аль зайгаар тэмцээн зохиож тамирчдын амжилтыг бүртгэвэл оновчтой болох тухай статистик дүгнэлт гаргана.

Агуулга

1 Өгөгдөл

Хөнгөн атлетикийн гүйлтийн төрлөөрх 54 улс орны үндэсний рекорд амжилтыг 2005 оны байдлаар, м/с-ээр илэрхийлсэн өгөгдөл авч үзнэ.

```
X <- read.csv(file = "data.csv", check.names = FALSE)
```

Өгөгдлийг файлаас ачаалж, датафрейм хэлбэртэй объект болгон X гэсэн нэрээр ажлын огторгуйд хадгалав. Датафреймын эхний хэдэн мөрийг Хүснэгт 1 дээр харуулав.

Датафреймын эхний багана бидэнд шаардлагагүй тул түүнийг зайлуулна. Бас баганын нэрд орсон кирилл үсгээс болж RMarkdown ашиглаж тайлан бэлдэх үед гарах анхааруулгаас зайлсхийхийн тулд баганын нэрсийг өөрчлөв.

```
X <- X[-1]
colnames(X) <- c(100, 200, 400, 800, 1500, 5000, 10000)
```

	Улс	100 м	200 м	400 м	800 м	1500 м	5000 м	10000 м
1	Австрали	10.070	9.970	9.013	7.663	7.082	6.445	6.054
2	Австри	9.852	9.780	8.734	7.533	6.983	6.285	6.013
3	АНУ	10.225	10.352	9.264	7.797	7.225	6.425	6.121
4	Аргентин	9.775	9.818	8.662	7.533	6.793	6.252	6.028
5	Бельги	9.862	9.906	8.885	7.707	7.003	6.495	6.203
6	Бермуд	9.737	9.852	8.838	7.449	6.757	5.692	5.466

Хүснэгт 1: Өгөгдлийн эхний 6 мөр

2 Корреляцийн шинжилгээ

100, 200, 400, 800, 1500, 5000, 10000 хувьсагчдын хувьд түүврийн корреляцийн матрицыг дараах байдлаар олж болно.

```
R <- cor(X)
```

Ийнхүү олсон түүврийн корреляцийн матриц

$$R = \begin{pmatrix} 1.00 & 0.91 & 0.79 & 0.70 & 0.75 & 0.73 & 0.70 \\ 0.91 & 1.00 & 0.84 & 0.78 & 0.79 & 0.75 & 0.73 \\ 0.79 & 0.84 & 1.00 & 0.75 & 0.74 & 0.76 & 0.74 \\ 0.70 & 0.78 & 0.75 & 1.00 & 0.88 & 0.85 & 0.83 \\ 0.75 & 0.79 & 0.74 & 0.88 & 1.00 & 0.90 & 0.89 \\ 0.73 & 0.75 & 0.76 & 0.85 & 0.90 & 1.00 & 0.99 \\ 0.70 & 0.73 & 0.74 & 0.83 & 0.89 & 0.99 & 1.00 \end{pmatrix}$$

байна. Матрицаас 100, 200, 400, 800, 1500, 5000, 10000 хувьсагчид өөр хоорондоо маш өндөр, шууд хамааралтай болох нь харагдана. Иймд эдгээр хувьсагчдын холбоо хамаарлыг цааш үргэлжлүүлэн судалбал зохино.

3 Гол хэсгийн шинжилгээ

Корреляцийн шинжилгээгээр 100, 200, 400, 800, 1500, 5000, 10000 хувьсагчид хоорондоо хүчтэй хамааралтай гэдгийг тогтоосон. Иймд тэдгээр хувьсагчдын холбоо хамаарлыг цааш нь нарийвчлан шинжилж болох юм.

Эхлээд дээрх нэр бүхий 7 хувьсагч угтаа хэчнээн хүчин зүйлээр тодорхойлогдож буйг олж тогтооно. Үүнд гол хэсгийн шинжилгээ [4, §3] хэрэг болно. Энэхүү шинжилгээг FactoMineR [3] болон factoextra [1] багц ашиглаж хийх боломжтой. Эдгээр багцууд нь гол хэсгийн шинжилгээний олон талын үр дүнг хялбар тушаалаар гаргаж харуулдаг. Мөн тус багцуудыг ашиглахаас өмнө тэдгээрийг хэрхэн ашиглах тухай, жишээ бүхий зөвлөмж [2] үзвэл зохимжтой.

```
pca <- FactoMineR::PCA(X = X, scale.unit = TRUE, graph = FALSE, ncp = 7)
```

Шинжилгээнд оруулсан хувьсагчид хэмжээсийн төрөл, масштабын хувьд бүгд ижил, тодруулбал м/с гэсэн нэгжээр илэрхийлэгдсэн тул мастабын ялгаатай байдал арилгах буюу стандарт хувиргалт хийх шаардлагагүй юм. Иймд тус функцийг scale.unit = FALSE аргументтайгаар дуудаж болно. Харин хувьсагчдын масштаб ялгаатай буюу өөр хэмжээсээр хэмжсэн өгөгдлийн хувьд scale.unit = TRUE аргумент дамжуулна. Ийм утга дамжуулах буюу хувьсагч дээр стандарт хувиргалт хийх нь бүх хувьсагчдыг ижил тэгш эрхтэй болгож буй явдал юм.

Энэхүү өгөгдөл дэх хувьсагчдын хувьд ойрын зайд илүү өндөр хурд гарах бол харин хол зайд бага дундаж хурд гарах тул хувьсагчдын масштаб ялгаатай гэж үзлээ. Иймд хувьсагчид дээр стандарт хувиргалт хийх буюу PCA() функцийг scale.unit = TRUE гэсэн байдлаар дуудах нь зүйтэй.

Ийнхүү хувьсагчид дээр стандарт хувиргалт хийх буюу хувьсагчдыг тэгш эрхтэй байдлаар авах уу гэдэг талаар хоёронтаа бодож байж эцсийн шийдвэрт хүрлээ.

Түүнчлэн шинжилгээнд 7 хувьсагч оруулсан тул 7 гол хэсэг олдох бөгөөд эдгээр бүгдийг эцсийн үр дүнд тусгуулахын тулд ncp аргументаар 7 гэсэн утга дамжууллаа.

3.1 Ялгаж авах гол хэсгийн тоо

Хэчнээн хүчин зүйл буюу гол хэсэг шаардлагатайг $\lambda_1, \ldots, \lambda_p$ хувийн утгууд буюу гол хэсгүүдийн дисперсүүдийн тусламжтай зохиогдох

$$\psi_q = \frac{\lambda_1 + \ldots + \lambda_q}{\lambda_1 + \ldots + \lambda_p} \cdot 100\% \qquad q = 1, \ldots, p$$

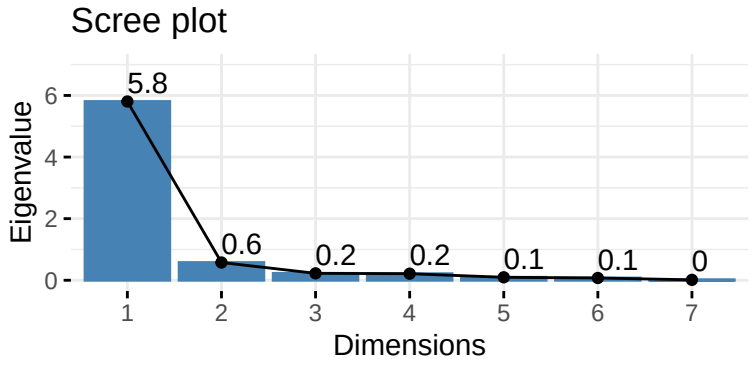
харьцаа дээр үндэслэж олдог. Тэдгээр хувийн утгууд болон энэхүү статистикийг FactoMineR::PCA() функцийн буцаах утгын eig элемент дэх матрицын эхний болон сүүлийн баганууддаас харж болно. Үүнийг хэвлэж харахын тулд pca$eig хэлбэртэй тушаал өгнө.

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	5.80375	82.91078	82.91078
comp 2	0.57592	8.22744	91.13822
comp 3	0.22456	3.20804	94.34626
comp 4	0.21307	3.04382	97.39008
comp 5	0.09625	1.37494	98.76502
comp 6	0.07401	1.05727	99.82229
comp 7	0.01244	0.17771	100.00000

Хүснэгт 2: Хувийн утгууд

Эндээс гарах үр дүнг Хүснэгт 2 дээр харуулав. Хүснэгтийн нэг дүгээр баганад буй хувийн утгуудыг диаграммаар харуулбал зохимжтой. Үүний тулд дараах байдалтай тушаал өгнө. Гарсан үр дүнг Зураг 1 дээрээс харна уу.

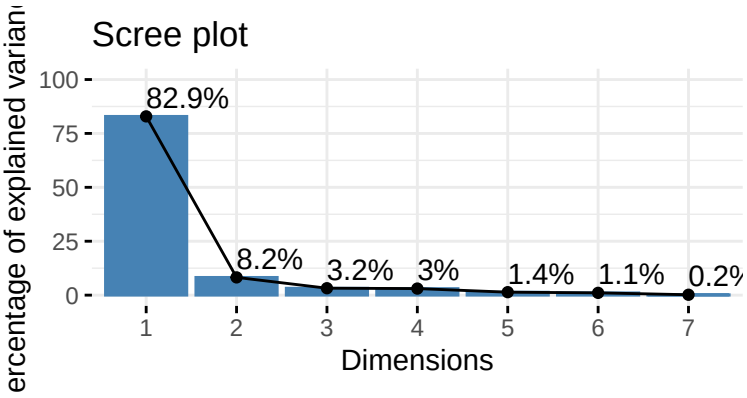
```
factoextra::fviz_screeplot(pca, choice = "eigenvalue", addlabels = TRUE, ylim = c(0,7))
```



Зураг 1: Хувийн утгууд

Харин хувийн утга тус бүр нь нийт хувийн утгуудын нийлбэрийн хэчнээн хувийг эзэлж буйг өөрөөр хэлбэл гол хэсэг тус бүр нийт дисперсийн хэдэн хувийг илэрхийлж буйг хоёр дугаар багана дахь утгууд харуулах бөгөөд үүгээр диаграмм байгуулахын тулд дараах хэлбэртэй тушаал өгнө. Байгуулсан диаграммыг Зураг 2 дээрээс харна уу.

```
factoextra::fviz_screeplot(pca, choice = "variance", addlabels = TRUE, ylim = c(0,100))
```



Зураг 2: Нийт хувийн утгуудын дотор хувийн утга тус бүрийн эзлэх хувь

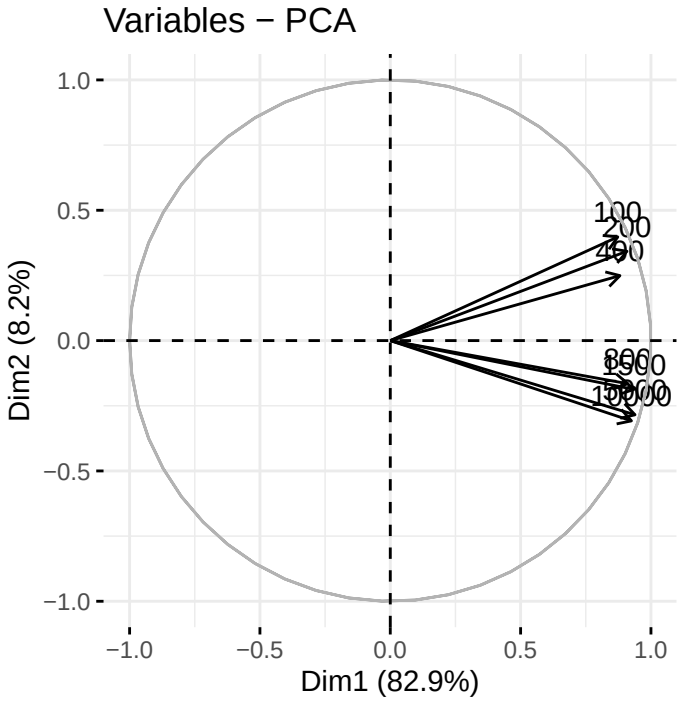
Эцэст нь гурав дугаар баганыг харвал $\psi_2 \approx 92.08$ буюу эхний хоёр гол хэсэг хамтдаа нийлээд нийт дисперсийн 90 гаруй хувийг эзэлж байна. Иймд дээрх 7 хувьсагчид агуулагдах мэдээлэл үндсэндээ хоёр хүчин зүйлээр тодорхойлогдоно хэмээн дүгнэж болно.

3.2 Гол хэсгүүдийг тайлбарлах нь

3.2.1 Анхны хувьсагчид болон гол хэсгүүд хоорондын корреляц

Гарсан үр дүнг тайлбарлахын тулд анхны хувьсагчид болон гол хэсгүүд хоорондын корреляцаар байгуулсан диаграмм байгуулсанаа Зураг 3 дээр харуулав. Тус диаграммыг байгуулахын тулд дараах тушаал өгнө.

```
factoextra::fviz_pca_var(pca)
```



Зураг 3: Эхний хоёр гол хэсэг болон хувьсагч хоорондын корреляц

Энэхүү корреляцийн диаграмм нь зөвхөн scale.unit = TRUE үед гарна.

Гол хэсгийн шинжилгээний корреляцийн диаграммыг харвал бүх хувьсагчид эхний гол хэсэгтэй маш өндөр, шууд хамааралтай байна. Харин хоёр дугаар гол хэсгийн зүгээс харвал ойрын зайн төрлүүд тус гол хэсэгтэй шууд хамааралтай, бусад төрлүүд урвуу хамааралтай ажээ. Бас энэхүү холбоо хамаарлаараа илэрхий хоёр тусдаа бүлэг үүсгэж байна. Эхний бүлэгт 100, 200, 400 метрийн төрлүүд харин нөгөө бүлэгт бусад төрлүүд багтаж байна. Ийнхүү ойрын болон холын зайн тус бүр нэг төрөлд тамирчдыг уралдуулбал зохимжтой гэсэн статистик дүгнэлт гарлаа.

3.2.2 Гол хэсэг дээрх хувьсагч тус бүрийн оролцоо

Гол хэсэг дээрх хувьсагч тус бүрийн оролцоо нь FactoMineR::PCA() функцийн буцаах утгын var элементийн contrib утга дотор байдаг. Тус функцийн утгыг pca гэсэн нэртэй хувьсагч буюу объект байдлаар ажлын огторгуйд хадгалсан тул уг оролцоог print(pcavarcontrib) тушаалаар хэвлэж харах боломжтой. Үүнийг Хүснэгт 3 дээр байрлууллаа.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7
100	13.12	27.46	13.18	12.06	0.23	33.19	0.74
200	14.20	20.39	5.15	1.14	4.83	53.47	0.81
400	13.37	10.79	68.11	1.25	5.87	0.59	0.03
800	14.31	4.75	1.02	55.94	14.21	9.46	0.30
1500	15.06	6.08	10.42	1.67	66.05	0.65	0.06
5000	15.19	14.06	0.78	12.11	3.34	0.14	54.38
10000	14.75	16.47	1.33	15.83	5.46	2.50	43.67

Хүснэгт 3: Гол хэсэг дээрх хувьсагч тус бүрийн оролцоо

Хүснэгтийн эхний баганыг харвал хувьсагчдад харгалзах утгууд ойролцоо байна. Энэ нь нэг дүгээр гол хэсэг дээрх хувьсагчдын оролцоо жигдхэн байна гэсэн үг юм. Хоёр дугаар гол хэсгийн хувьд 800 болон 1500 метрийн төрлүүдээс бусдынх оролцоо өндөр байна. Ийнхүү ойрын болон холын зайн төрлүүдийн ялгаа нь хоёр дугаар хүчин зүйлээр тайлбарлагдаж байна гэсэн дүгнэлт гарч буй нь корреляцийн диаграммаас ажигласантай тохирч байна. Харин 800 ба 1500 метр гэсэн дундын зайн ангиллууд ач холбогдол нь маш бага байсан 4 ба 5 дугаар гол хэсэг тус бүртэй уялдаж байна. Мөн энд 400 метрийн зай нь 3 дугаар гол хэсэгтэй илүү уялдаатай байгаа тэмдэглэх нь зүйтэй. Хэрэв энэхүү хоёр статистик дүгнэлт дээр ач холбогдол өгнө гэвэл дундын зайн ангилал нэмж болох юм.

4 Дүгнэлт

Хөнгөн атлетикийн гүйлтийн төрлөөрх 54 улс орны үндэсний рекорд амжилтыг 2005 оны байдлаар, м/с-ээр илэрхийлсэн өгөгдөл дээрх гол хэсгийн шинжилгээнээс дараах үр дүн гарлаа.

1. Гүйлтийн төрлийг уралдах зайгаар нь 7 ангилалд хуваасан нь хэт ихдэж байна.
2. Ойрын эсвэл холын зай гэсэн хоёр ангид хуваавал судалгаанд ашигласан өгөгдөл дэх нийт мэдээллийн 90 гаруй хувийг хамрахаар байна.
3. Хэрэв дундын зай гэсэн анги нэмбэл гол хэсгийн шинжилгээгээр олж тогтоосон 5 хүчин зүйлийн нөлөөг харгалзан үзсэн явдал болох бөгөөд энэ нь нийт өгөгдөл дэх дисперсийн 1 гаруйхан хувийг үл тооцно гэсэн үг юм.

Ийнхүү хөнгөн атлетикийн гүйлтийн төрөлд тамирчдын уралдах зай нь ойр, дунд, хол гэсэн гурван шат бүхий ангилалтай байвал зохимжтой гэсэн статистик дүгнэлт гарав.

Ашигласан материал

[1] Extract and visualize the results of multivariate data analyses. URL: https://rpkgs.datanovia.com/factoextra/index.html.

[2] FactoMineR and factoextra : Principal Component Analysis Visualization - R software and data mining - Easy Guides - Wiki - STHDA. en. URL: http://www.sthda.com/english/wiki/wiki.php?id_contents=7851 (urlseen 15/04/2022).
[3] Sébastien Lê, Julie Josse and François Husson. "FactoMineR: An R Package for Multivariate Analysis". in: Journal of Statistical Software 25.1 (2008), pages 1–18. DOI: 10.18637/jss.v025.i01. URL: https://www.jstatsoft.org/index.php/jss/article/view/v025i01.
[4] Г.Махгал Ш.Мөнгөнсүх. Олон хэмжээст өгөгдлийн статистик шинжилгээ. 2017. ISBN: 9789997816481.