Олон хэмжээст өгөгдлийн статистик шинжилгээ хичээлийн нэмэлт жишээ

Монгол эмэгтэйчүүдийн биеийн хэмжээний зарим хэмжээс

Г.Махгал

2019 оны 2 сарын 2

Агуулга

1 Өгөгдөл

Өгөгдлийг файлаас уншиж, орхигдсон утгуудыг зайлуулаад датафрейм хэлбэртэй объект болгон X гэсэн нэрээр ажлын огторгуйд хадгалав. Датафреймын эхний хэдэн мөрийг хүснэгт 1 дээр харуулав.

	Нас	Өндөр	Цээжний тойрог	Бүсэлхийн тойрог	Өгзөгний тойрог
1	30	165.7	105.0	93.0	111.5
2	23	152.9	87.0	72.0	91.0
3	24	151.2	81.3	73.0	89.0
4	29	160.0	90.5	64.5	92.0
5	48	150.8	86.0	74.0	96.0
6	32	163.0	90.0	75.5	97.0

Хүснэгт 1: Өгөгдлийн эхний 6 мөр

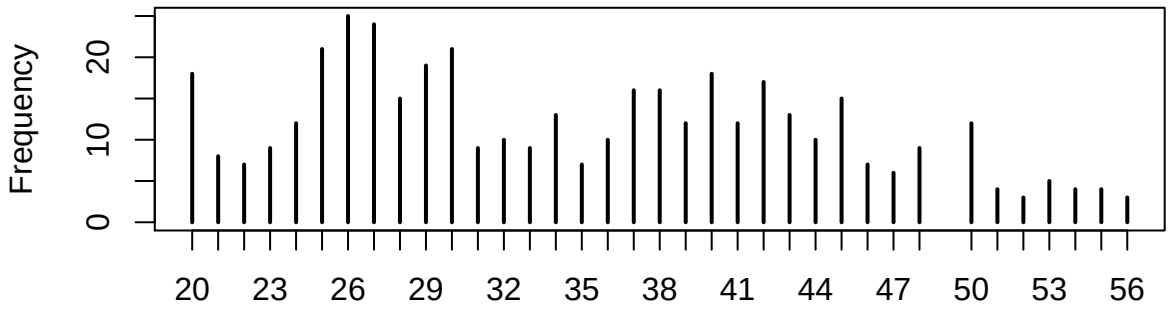
X датафрейм 423 мөр болон age, height, chest_line, waist_line, hip_line гэсэн 5 баганатай.

X$age дискрет хувьсагчийн нягтыг харуулсан диаграммыг

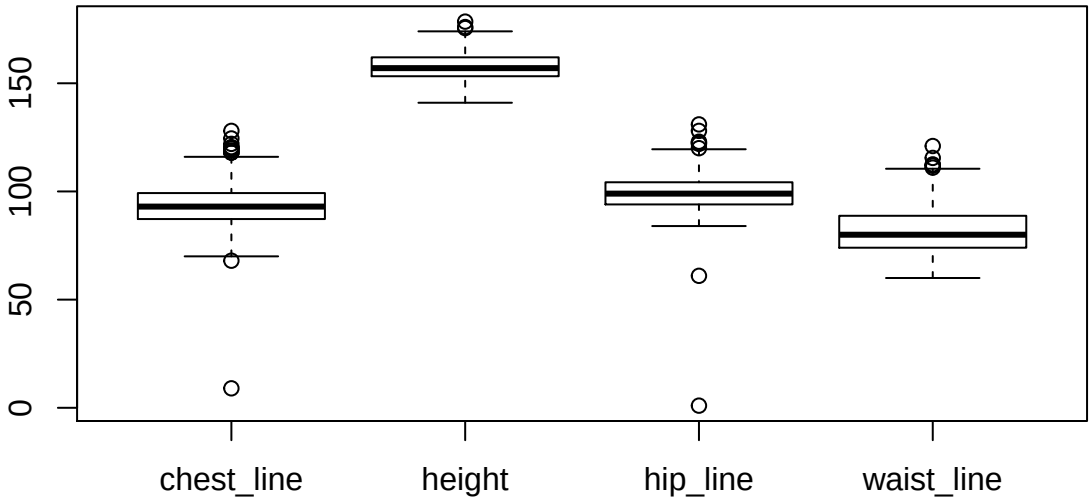
```
plot(table(X$age), type = "h", ylab = "Frequency")
```

тушаалаар байгуулсаныг зураг 1 дээр харуулав. Бусад дөрвөн хувьсагчдын хувьд хайрцган диаграмм байгуулъя. Үүний тулд дараах байдлаар өргөн бүтэцтэй өгөгдлөө урт бүтцэд шилжүүлээд томьёололд тулгуурласан boxplot() функц ашиглана.

```
X.long <- tidyr::gather(data = X, key = category, value = value, height, chest_line, waist_line, hip_line)
boxplot(formula = value ~ category, data = X.long)
```



Зураг 1: Нас хувьсагчийн утгуудын давтамж



Зураг 2: Өндөр болон тойрог хэмжээний хувьсагчдын хайрцган диаграмм

Зураг 2 дээрх хайрцган диаграммыг харвал chest_line болон hip_line хувьсагчид алдаатай утга агуулжээ. Эдгээрийг дараах байдлаар шүүж цэвэрлэж болно.

```
X <- subset(x = X, subset = chest_line > 50 & hip_line > 50)
```

Ингээд өмнөх хайрцган диаграммыг ахин байгуулсаныг зураг 3 дээрээс харна уу.

Ингэснээр түүврийн хэмжээ 421 боллоо.

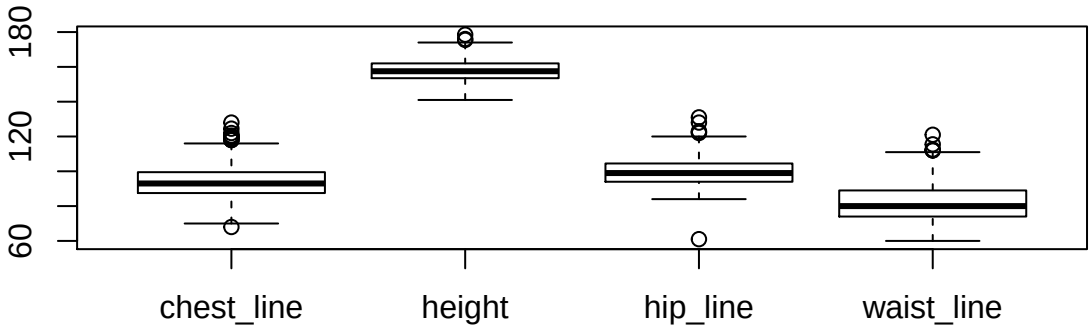
2 Олон хэмжээст хэвийн тархалт

height, chest_line, waist_line, hip_line хувьсагчдыг олон хэмжээст хэвийн тархалттай гэж тооцъё.

2.1 Тархалтын параметр

height, chest_line, waist_line, hip_line хувьсагчдын түүврийн дундаж утгын вектор

$$m = (157.89, 94.14, 81.61, 99.66)^T$$



Зураг 3: Алдаатай утгуудыг зайлуулсаны дараах өндөр болон тойрог хэмжээний хувьсагчдын хайрцган диаграмм

харин түүврийн ковариацийн матриц

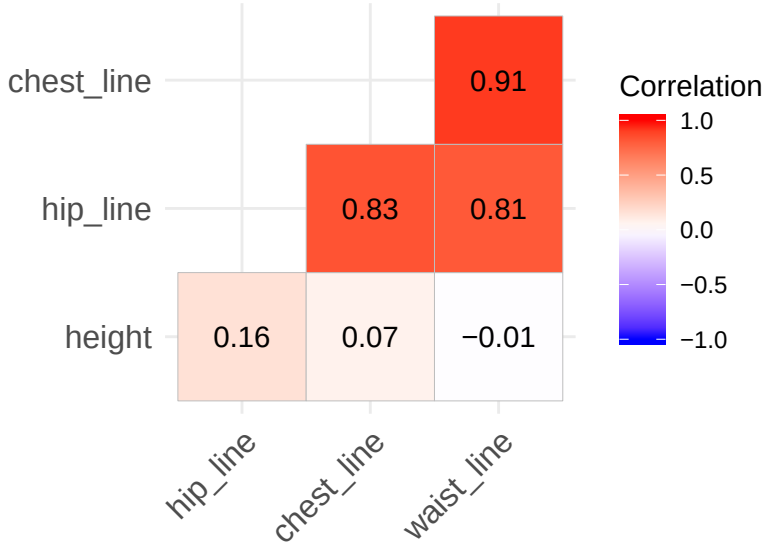
$$S = \begin{pmatrix} 37.77 & 4.31 & -0.81 & 7.66 \\ 4.31 & 87.71 & 88.33 & 61.64 \\ -0.81 & 88.33 & 107.87 & 67.14 \\ 7.66 & 61.64 & 67.14 & 63.38 \end{pmatrix}$$

болон корреляцийн матриц

$$R = \begin{pmatrix} 1.00 & 0.07 & -0.01 & 0.16 \\ 0.07 & 1.00 & 0.91 & 0.83 \\ -0.01 & 0.91 & 1.00 & 0.81 \\ 0.16 & 0.83 & 0.81 & 1.00 \end{pmatrix}$$

байна. Мөн корреляцийн матрицийг диаграммаар харуулж болдог бөгөөд үүний тулд

```
ggcorrplot::ggcorrplot(
  corr = R,
  legend.title = "Correlation", lab = TRUE, hc.order = TRUE, type = "lower"
)
```



Зураг 4: Корреляцийн матриц диаграммаар

хэлбэртэй тушаал өгч болно. Байгуулсан диаграммыг зураг 4 дээр харуулав.

2.2 Ковариацийн матрицийн тухай таамаглал

Корреляцийн матрицын бүтцээс chest_line, waist_line, hip_line хувьсагчдыг л хамтад нь авч судалбал зохих буюу age санамсаргүй хувьсагч болон height, chest_line, waist_line, hip_line санамсаргүй вектор хамааралгүй гэсэн таамаглал гарч байна. Өөрөөр хэлбэл [1, мөрдлөгөө 2.1] ёсоор height, chest_line, waist_line, hip_line хувьсагчдын ковариацийн матрицийг

$$\Sigma_0 = \begin{pmatrix} 37.77 & 0.00 & 0.00 & 0.00 \\ 0.00 & 87.71 & 88.33 & 61.64 \\ 0.00 & 88.33 & 107.87 & 67.14 \\ 0.00 & 61.64 & 67.14 & 63.38 \end{pmatrix}$$

хэлбэртэй гэсэн $H_0 : \Sigma = \Sigma_0$ таамаглал дэвшүүлж байна. Энэ нь [1, §2.11.2] дэх 3 дугаар таамаглал юм.

> Энэ таамаглал нь биеийн өндрийн хэмжээг цээж, бүсэлхий, өгзөгний тойрогт ямар ч нөлөөгүй гэсэн утга агуулгатай юм.

Таамаглалыг 0.05 итгэх түвшинд шалгахын тулд 5 дугаар лекц дээр үзсэн R хэл дээрх кодыг авч ашиглая. Шинжүүрийн статистикийн туршилтын утга 41.46 нь 0.05 итгэх түвшинд харгалзах шинжүүрийн няцаах утга 18.31-аас их байгаа тул H_0 таамаглалыг няцаана.

> Иймд биеийн өндрийн хэмжээг цээж, бүсэлхий, өгзөгний тойрогт нөлөөтэй гэж үзэх үндэслэлтэй ажээ.

2.3 Нөхцөлт тархалт

Санамсаргүй хувьсагчдыг

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \\ \begin{pmatrix} X_3 \\ X_4 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} \text{height} \\ \text{chest\_line} \\ \text{waist\_line} \\ \text{hip\_line} \end{pmatrix}$$

байдлаар авч

$$E(X_2|X_1 = x_1) = \mu_{2.1} + \Sigma_{21}\Sigma_{11}^{-1}x_1$$

нөхцөлт дундаж улмаар

$$\text{cov}(X_2|X_1 = x_1) = \Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

нөхцөлт ковариацийн матрицийг авч үзэх буюу өндрийн нөлөөг зайлуулсан үед тойрог хэмжээнүүдийн холбоо хамаарлыг судалъя.

2.3.1 Нөхцөлт дундаж

Нөхцөлт дундаж

$$E(X_2|X_1 = x_1) = \mu_{2.1} + \Sigma_{21}\Sigma_{11}^{-1}x_1 = \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1 + \Sigma_{21}\Sigma_{11}^{-1}x_1 = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)$$

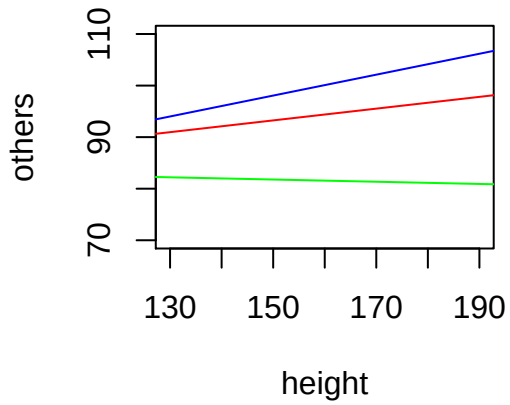
хэлбэртэй байхыг сануулъя.

$$E(X_2|X_1 = x_1) = \begin{pmatrix} 94.14 \\ 81.61 \\ 99.66 \end{pmatrix} + \begin{pmatrix} 4.31 \\ -0.81 \\ 7.66 \end{pmatrix} \begin{pmatrix} 37.77 \end{pmatrix}^{-1} (x_1 - 157.89)$$

Үүнийг цааш эмхэтгэвэл

$$E(X_2|X_1 = x_1) = \begin{pmatrix} 0.11 \cdot x_1 + 76.13 \\ -0.02 \cdot x_1 + 85.01 \\ 0.2 \cdot x_1 + 67.66 \end{pmatrix}$$

болно. Энд олдсон нөхцөлт математик дунджуудад харгалзах шулууныг зураг 5 дээр зурж харуулав.



Зураг 5: Цээж (улаан), бүсэлхий (ногоон), өгзөгний тойрог (цэнхэр) хувьсагчдын биеийн өндрийн нөхцөл дэх математик дундаж

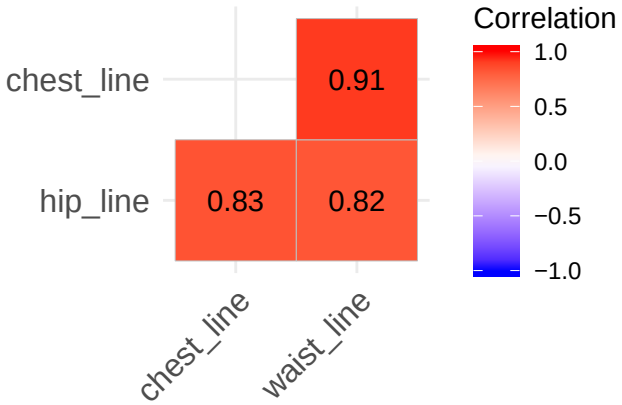
2.3.2 Нөхцөлт ковариац

height буюу өндөр хувьсагчийг бэхэлсэн үед үлдэх гурван хувьсагчийн тухайн ковариац болон корреляцийг олъё.

$$\begin{aligned}\operatorname{cov}(X_2|X_1=x_1) &= \begin{pmatrix} 87.71 & 88.33 & 61.64 \\ 88.33 & 107.87 & 67.14 \\ 61.64 & 67.14 & 63.38 \end{pmatrix} - \begin{pmatrix} 4.31 \\ -0.81 \\ 7.66 \end{pmatrix} \begin{pmatrix} 37.77 \end{pmatrix}^{-1} \begin{pmatrix} 4.31 & -0.81 & 7.66 \end{pmatrix} \\ &= \begin{pmatrix} 87.22 & 88.42 & 60.77 \\ 88.42 & 107.85 & 67.30 \\ 60.77 & 67.30 & 61.83 \end{pmatrix}\end{aligned}$$

$$\operatorname{cor}(X_2|X_1=x_1) = \begin{pmatrix} 1.00 & 0.91 & 0.83 \\ 0.91 & 1.00 & 0.82 \\ 0.83 & 0.82 & 1.00 \end{pmatrix}$$

Тухайн корреляцийн диаграммыг зураг 6 дээр харуулав.



Зураг 6: height хувьсагчийн нөлөөг зайлуулсан үеийн тухайн корреляцийн матриц диаграммаар

Үүнийг зураг 4 дээрх диаграммтай харьцуулбал hip_line болон waist_line буюу бэлхүүс болон өгзөгний тойрог гэсэн хувьсагчдын корреляц үл ялиг өссөн нь анзаарагдаж байна.

[1, код 2.3] аргаар олсон тухайн корреляцийг хүснэгт 2 дээр жагсаав.

	height	chest_line	waist_line	hip_line
height	1.00	0.11	-0.25	0.23
chest_line	0.11	1.00	0.72	0.33
waist_line	-0.25	0.72	1.00	0.30
hip_line	0.23	0.33	0.30	1.00

Хүснэгт 2: Тухайн корреляцийн коэффициент

Үүнийг тухайлбал цээжний тойргийн хэмжээ ба бүсэлхийн тойргийн хэмжээ хоёр бусад хоёр хувьсагч буюу өндөр болон өгзөгний тойргийн нөлөөг зайлуулсан үеийн хамаарлыг харуулсан корреляцийн коэффициент 0.72 байна гэж ойлгоно. Эндээс өндрийн хэмжээ ердийн корреляцтай харьцуулахад мэдэгдэхүйц өсч аль аль хувьсагчтай сул хамааралтай гэж харагдаж байгаа нь анхааруштай юм.

2.4 Дунджуудын ялгаврын тухай таамаглал

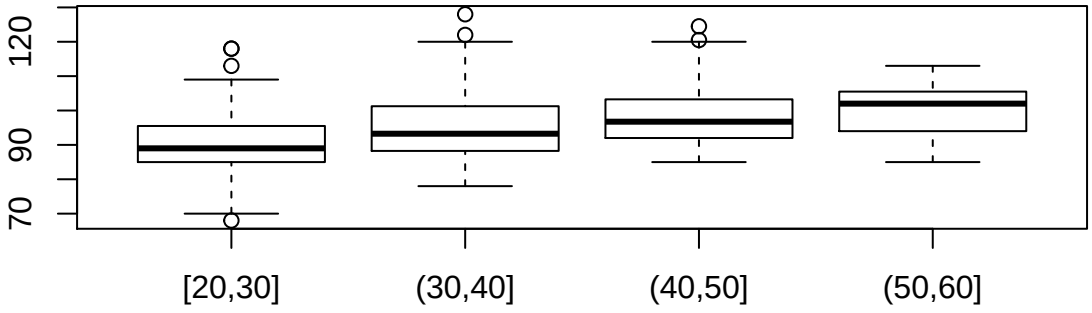
Өмнө авч үзсэн нөхцөлт дундажтай холбоотой үр дүнгээс харвал өндөр нэмэгдэхэд тойрог хэмжээнүүдэд тодорхой өөрчлөлт гарч буй мэт харагдаж байсан. Үүнийг батлахын тулд түүврийн дунджаас өндөр болон нам хүмүүсээс бүрдэх хоёр бүлэг chest_line, waist_line, hip_line вектор хувьсагчийн дунджаараа ялгаатай гэж харуулахад хангалттай. Үүний тулд [1, §2.11.2] хэсэг дэх таамаглал 7-г $\Delta\mu = 0$ үед шалгана. Шинжүүрийн статистикийн туршилтын утга 10.52, магадлалын утга 1.1062724×10^{-6} байна.

2.5 Олон хэмжээст дисперсийн шинжилгээ

Тойрог хэмжээнүүд наснаас хамаарах эсэхийг шалгаж үзье. Гэхдээ үүнд олон хэмжээст дисперсийн шинжилгээ ашиглана. Үүний тулд өгөгдлөө насаар нь [20,30], (30,40], (40,50], (50,60] гэж бүлэглэв. Энэ тохиолдолд тэг таамаглал нь [20,30], (30,40], (40,50], (50,60] бүлгүүд дээрх chest_line, waist_line, hip_line хувьсагчдаас тогтох санамсаргүй векторын математик дундаж адил тэнцүү гэсэн утгатай байна.

```
summary(manova(as.matrix(X[,3:5]) ~ age.group), test = "Wilks")
```

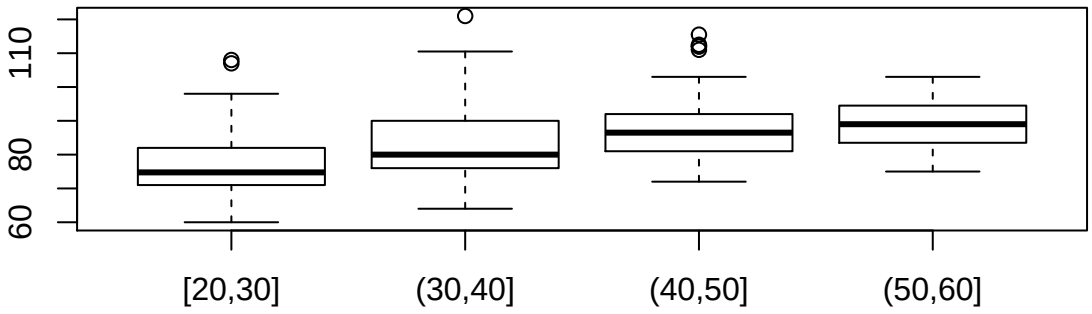
Дээрх байдлаар уг таамаглалыг шалгахад магадлалын утга нь $1.0437817 \times 10^{-19}$ гэж гарч байгаа тул тэг таамаглалыг няцаах бүрэн үндэслэлтэй гэж дүгнэнэ. Үнэхээр ч зураг 7, 8 болон 9 дээрх хайрцган диаграммуудыг харвал нас ахих тусам уг хувьсагчдын утга өсөх хандлагатай буюу нэг төрлийн бус ажээ.



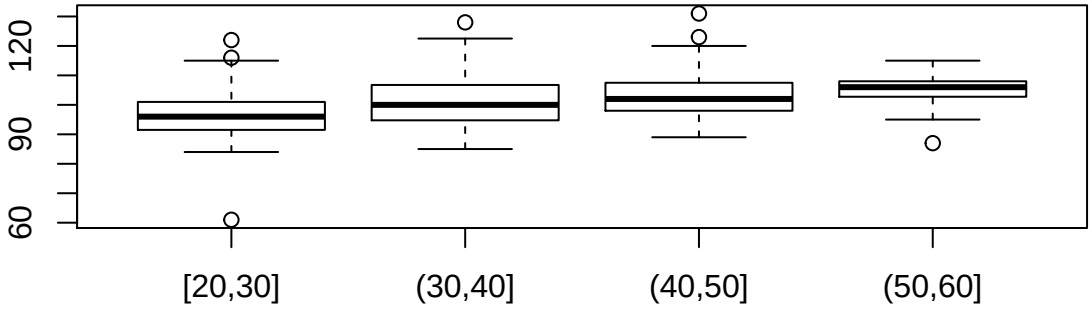
Зураг 7: Цээжний тойргийн хэмжээний тархалт насны бүлгээр

3 Гол хэсгийн шинжилгээ

height, chest_line, waist_line, hip_line хувьсагчид дээр гол хэсгийн шинжилгээ хийе.



Зураг 8: Бүсэлхийн тойргийн хэмжээний тархалт насны бүлгээр

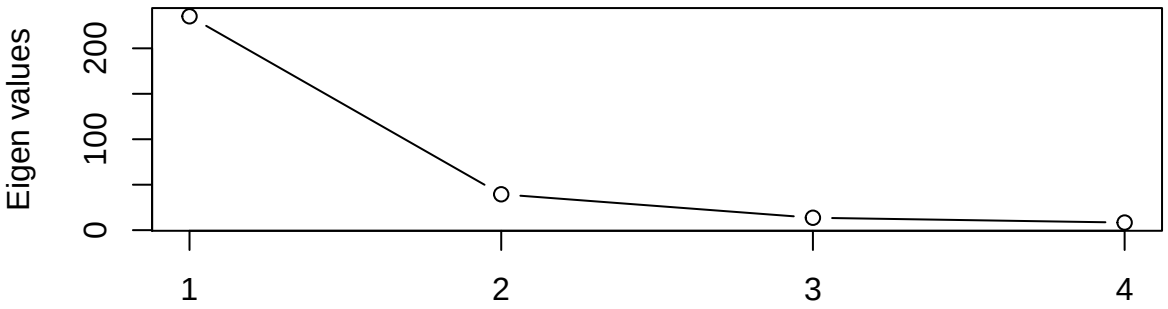


Зураг 9: Өгзөгний тойргийн хэмжээний тархалт насны бүлгээр

```
pca <- prcomp(x = X[,2:5])
```

Ковариацын матрицын хувийн утгууд ойролцоогоор 235.23, 39.47, 13.66, 8.38 бөгөөд зураг 10 дээр диаграммаар үзүүлэв.

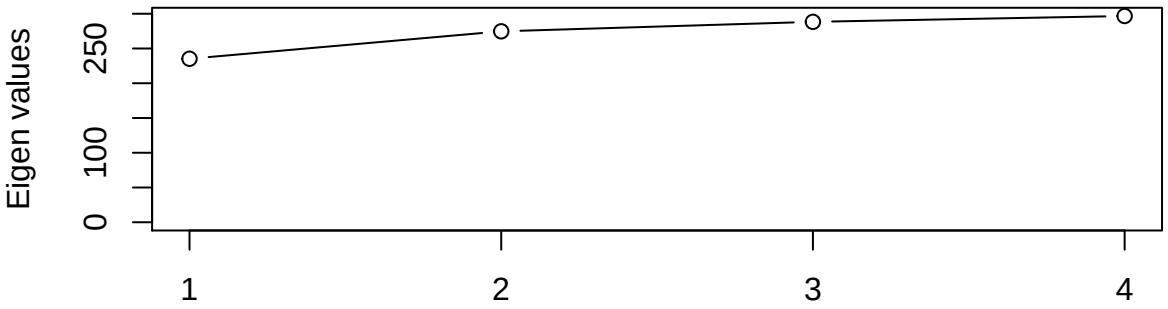
```
plot(x = 1:4, y = pca$sdev ** 2, type = "b", xlab = "", ylab = "Eigen values", xaxt = "n")
axis(side = 1, at=1:4, labels=1:4)
```



Зураг 10: Ковариацын матрицын хувийн утгууд

Харин хувийн утгуудын хуримтлагдах нийлбэрийг зураг 11 дээр диаграммаар харуулав.

Улмаар $\psi_q = \dfrac{\lambda_1 + \ldots + \lambda_q}{\lambda_1 + \ldots + \lambda_4}$ хэмжигдэхүүний утгыг $q = 1, \ldots, 4$ үед тооцоолон зууны хувиар илэрхийлбэл харгалзан 79.27, 92.57, 97.18, 100 үр дүн гарч байна. Иймд эхний хоёр гол хэсгийг авбал зохимжтой ажээ. Энэ тохиолдолд уг гол хэсгүүд анхны өгөгдөл дэх нийт дисперсийн 92.57 хувийг илэрхийлнэ.

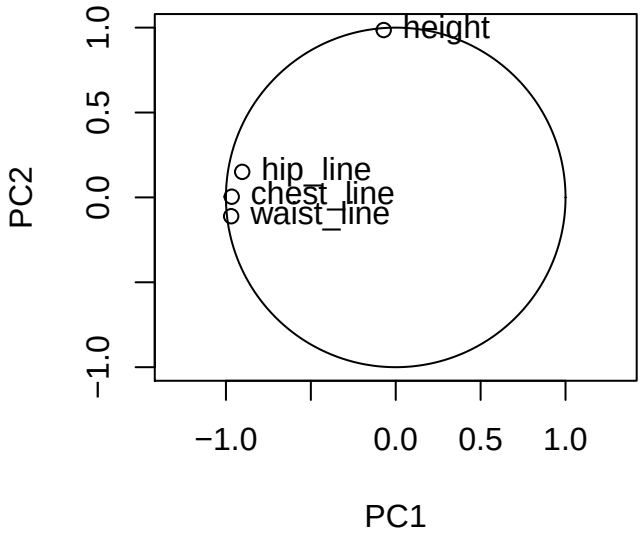


Зураг 11: Ковариацын матрицын хувийн утгуудын хуримтлагдах нийлбэр

Эцэст нь эхний хоёр гол хэсэг ба анхны дөрвөн хувьсагч хоорондын корреляц

$$\begin{pmatrix} -0.07 & 0.99 \\ -0.97 & 0.00 \\ -0.97 & -0.11 \\ -0.90 & 0.15 \end{pmatrix}$$

дээр үндэслэн байгуулсан диаграммыг зураг 12 дээр үзүүллээ.



Зураг 12: Хувьсагчид ба эхний хоёр гол хэсэг хоорондын корреляц

Диаграммаас тойрог хэмжээний гурван хувьсагч бүгд зөвхөн эхний гол хэсэгтэй өндөр холбоо хамааралтай бол харин өндөр гэсэн хувьсагч зөвхөн хоёр дахь гол хэсэгтэй л холбоо хамааралтай байгааг харж болно. Иймд энд цээж, бүсэлхий, өгзөгний тойрог гурвыг нийлүүлсэн нэг хүчин зүйл бас өндрийн хэмжээ гэсэн хувьсагчтай шууд холбогдох өөр нэг хүчин зүйл буюу нийтдээ хоёр үндсэн хүчин зүйл байна гэж дүгнэж болно.

Ашигласан материал

[1] Г.Махгал Ш.Мөнгөнсүх. Олон хэмжээст өгөгдлийн статистик шинжилгээ. 2017. ISBN: 9789997816481.