

# Los peces y el mercurio

David Núñez A01634928

2022-11-30

La contaminación por mercurio de peces en el agua dulce comestibles es una amenaza directa contra nuestra salud. Se llevó a cabo un estudio reciente en 53 lagos de Florida con el fin de examinar los factores que influían en el nivel de contaminación por mercurio. La descripción de los datos es la siguiente:

X1 = número de indentificación

X2 = nombre del lago

X3 = alcalinidad (mg/l de carbonato de calcio)

X4 = PH

X5 = calcio (mg/l)

X6 = clorofila (mg/l)

X7 = concentración media de mercurio (parte por millón) en el tejido muscular del grupo de peces estudiados en cada lago

X8 = número de peces estudiados en el lago

X9 = mínimo de la concentración de mercurio en cada grupo de peces

X10 = máximo de la concentración de mercurio en cada grupo de peces

X11 = estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible)

X12 = indicador de la edad de los peces (0: jóvenes; 1: maduros)

Para dar solución a este problema se utilizaron metodos estadisticos tales como distribución normal multivariada debido a que los atributos de los datos pudieran estar correlacionados y PCA (componentes principales) para obtener las combinaciones lineales de las de las variables que explican la varianza de los datos en pro de determinar su inferencia en la concentración de Mercurio en los peces.

## Normal multivariada

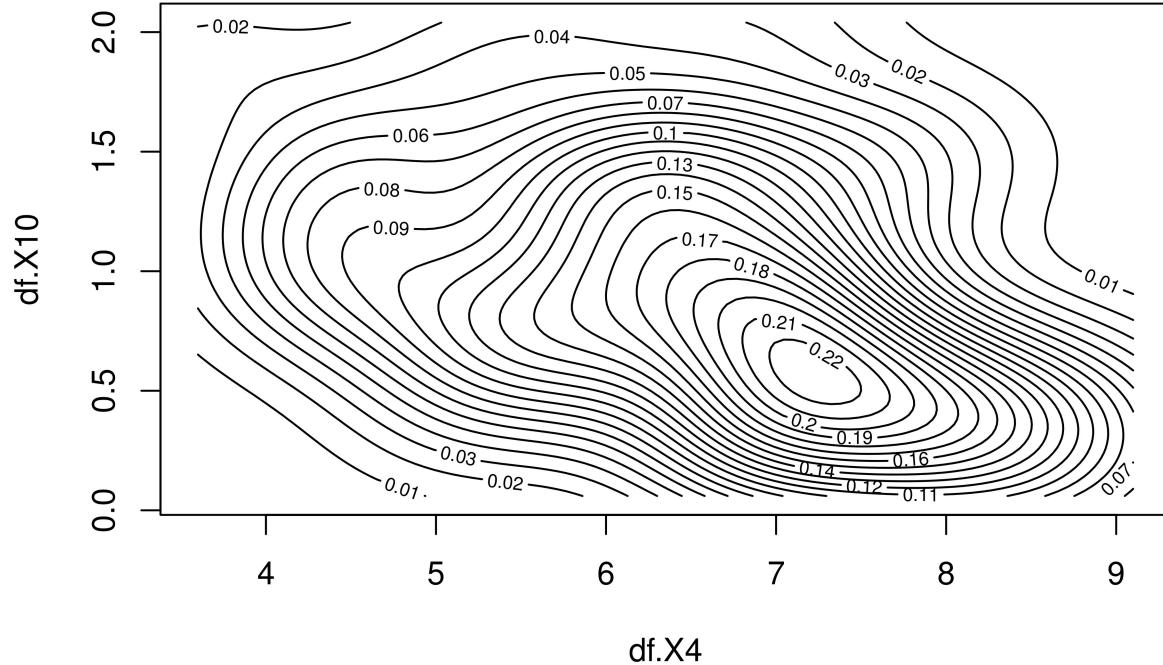
El test de Anderson-Darling se define como:

H0: es que sigue una distribución normal

H1: no sigue una distribución normal

```
##           Test  Variable Statistic   p value Normality
## 1 Anderson-Darling    X3      3.6725 <0.001      NO
## 2 Anderson-Darling    X4      0.3496  0.4611     YES
## 3 Anderson-Darling    X5      4.0510 <0.001      NO
## 4 Anderson-Darling    X6      5.4286 <0.001      NO
## 5 Anderson-Darling    X7      0.9253  0.0174      NO
## 6 Anderson-Darling    X8      8.6943 <0.001      NO
## 7 Anderson-Darling    X9      1.9770 <0.001      NO
## 8 Anderson-Darling   X10      0.6585  0.081      YES
## 9 Anderson-Darling   X11      1.0469  0.0086      NO
```

Tomando en cuenta las variables numericas se muestra normalidad bivariada, debido a que tenemos dos variables normales X4 y X10, debido a que p es mayor a 0.05 lo tanto se cumple H0 sigue una distribución normal.



Se observa sesgo a la izquierda en la media de X4 y sesgo a la derecha de la media de X10

```
## multivariateOutliers: NULL
```

NO se detectan valores atípicos en la serie de datos. Por lo tanto se prosigue a realizar análisis de componentes principales

## PCA

Para obtener los componenetes principales se necesita calcular una matriz de covarianzas y correlaciones para las variables numericas de nuestro dataframme.

```
##          X3        X4        X5        X6        X7        X8
## X3 1459.509456 35.3997134 793.065711 562.193324 -7.73773984 3.36556604
## X4   35.399713  1.6601016 18.540018 24.159971 -0.25283491 -0.20522496
## X5   793.065711 18.5400181 621.633266 314.949198 -3.40693687 -19.07703193
## X6   562.193324 24.1599710 314.949198 949.645668 -5.16408563 -3.11828737
## X7   -7.737740 -0.2528349 -3.406937 -5.164086  0.11630530  0.23074020
## X8    3.365566 -0.2052250 -19.077032 -3.118287  0.23074020 73.28519594
## X9   -4.544071 -0.1580980 -1.876788 -2.793997  0.07159176 -0.15825835
## X10  -12.062062 -0.3711680 -5.309432 -7.802021  0.16305729  0.71993106
## X11   -8.126195 -0.2674692 -3.922122 -5.286440  0.11080733  0.07481495
```

```

##          X9          X10          X11
## X3 -4.54407112 -12.06206241 -8.12619485
## X4 -0.15809797 -0.37116800 -0.26746916
## X5 -1.87678810 -5.30943179 -3.92212155
## X6 -2.79399673 -7.80202068 -5.28644013
## X7  0.07159176  0.16305729  0.11080733
## X8 -0.15825835  0.71993106  0.07481495
## X9  0.05125958  0.09046049  0.07048523
## X10 0.09046049  0.27253295  0.15203327
## X11 0.07048523  0.15203327  0.11473759

##          X3          X4          X5          X6          X7          X8
## X3  1.00000000  0.71916568  0.83260419  0.47753085 -0.59389671  0.01029074
## X4  0.71916568  1.00000000  0.57713272  0.60848276 -0.57540012 -0.01860607
## X5  0.83260419  0.57713272  1.00000000  0.40991385 -0.40067958 -0.08937901
## X6  0.47753085  0.60848276  0.40991385  1.00000000 -0.49137481 -0.01182027
## X7 -0.59389671 -0.57540012 -0.40067958 -0.49137481  1.00000000  0.07903426
## X8  0.01029074 -0.01860607 -0.08937901 -0.01182027  0.07903426  1.00000000
## X9 -0.52535654 -0.54196524 -0.33247623 -0.40045856  0.92720506 -0.08165278
## X10 -0.60479558 -0.55181523 -0.40791663 -0.48497215  0.91586397  0.16109174
## X11 -0.62795845 -0.61284905 -0.46440947 -0.50644193  0.95921481  0.02580046

##          X9          X10          X11
## X3 -0.52535654 -0.6047956 -0.62795845
## X4 -0.54196524 -0.5518152 -0.61284905
## X5 -0.33247623 -0.4079166 -0.46440947
## X6 -0.40045856 -0.4849721 -0.50644193
## X7  0.92720506  0.9158640  0.95921481
## X8 -0.08165278  0.1610917  0.02580046
## X9  1.00000000  0.7653532  0.91908939
## X10 0.76535319  1.0000000  0.85975810
## X11 0.91908939  0.8597581  1.00000000

```

Una vez calculados se obtienen los eigen valores y vectores propios de cada matriz.

```

## eigen() decomposition
## $values
## [1] 2.256459e+03 6.326276e+02 1.473932e+02 6.886530e+01 6.536708e-01
## [6] 2.529759e-01 3.020380e-02 4.651955e-03 1.946687e-03
##
## $vectors
##           [,1]           [,2]           [,3]           [,4]           [,5]
## [1,]  0.770052707 -0.3595628867  0.512208576 -1.212294e-01 -0.023208046
## [2,]  0.020607444  0.0064784700  0.013447171 -7.835184e-03  0.989042929
## [3,]  0.459104487 -0.2605992620 -0.824602008  2.030053e-01  0.006231571
## [4,]  0.442395277  0.8959627956 -0.034881281  8.630735e-03 -0.015103512
## [5,] -0.004349946 -0.0015154350 -0.006280700  6.265597e-03 -0.070456679
## [6,] -0.003461124  0.0017240484  0.236858719  9.714347e-01  0.005749934
## [7,] -0.002482186 -0.0006039179 -0.004911710 -6.851298e-05 -0.064182335
## [8,] -0.006732316 -0.0020103775 -0.009275875  1.488313e-02 -0.073080876
## [9,] -0.004611180 -0.0012562489 -0.004970505  3.214838e-03 -0.080815413
##           [,6]           [,7]           [,8]           [,9]
## [1,]  0.011728199  0.0017253286  0.0001564783  6.253514e-05
## [2,]  0.140012239 -0.0375235348 -0.0004638668 -8.577476e-03

```

```

## [3,] -0.009417996 -0.0016136975 -0.0011527025 -8.392504e-05
## [4,] 0.004138600 0.0006108642 -0.0003813870 3.630229e-04
## [5,] 0.472098416 -0.2823081524 0.3073272995 7.732471e-01
## [6,] -0.010565321 -0.0070248213 0.0006326070 -1.805388e-03
## [7,] 0.295149427 -0.4661481465 0.5865444157 -5.894136e-01
## [8,] 0.693632901 0.6928389806 0.0108480697 -1.817475e-01
## [9,] 0.434654887 -0.4706578821 -0.7492634949 -1.468724e-01

## eigen() decomposition
## $values
## [1] 5.34590819 1.22090789 1.04253153 0.66786333 0.33571266 0.20893778 0.10725403
## [8] 0.05203127 0.01885332
##
## $vectors
## [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## [1,] -0.35136146 -0.40301855 -0.07586402 0.30359419 0.03194121 0.284360283
## [2,] -0.33907420 -0.29786166 -0.07470140 -0.23236707 -0.82623084 0.054271109
## [3,] -0.28306469 -0.56943030 0.02991336 0.37427137 0.32816132 -0.298278080
## [4,] -0.28126962 -0.21524882 -0.06147214 -0.83056128 0.39488490 -0.099142969
## [5,] 0.39890941 -0.32518645 -0.05648045 -0.04980219 -0.06539303 0.004765464
## [6,] 0.02398876 0.06261499 -0.96994179 0.05149024 0.09004998 0.149954574
## [7,] 0.36905050 -0.37647100 0.11743644 -0.11401063 0.10565624 0.489107573
## [8,] 0.37957032 -0.24428857 -0.16175615 -0.02767633 -0.16523448 -0.711214479
## [9,] 0.40293860 -0.25922456 0.00756517 -0.07091614 -0.04298253 0.223233955
##      [,7]      [,8]      [,9]
## [1,] 0.72620919 -0.082971700 0.007161703
## [2,] -0.22348526 0.009782475 -0.032988603
## [3,] -0.48766992 0.140957430 -0.017292418
## [4,] 0.11144724 0.043959526 0.028777382
## [5,] 0.01398475 -0.053416125 0.849768758
## [6,] -0.14013431 -0.011952152 -0.041106334
## [7,] -0.22360542 -0.528271290 -0.340326567
## [8,] 0.30736177 -0.211913074 -0.311145559
## [9,] 0.09015694 0.802648566 -0.247594211

```

Estos valores serán útiles para determinar la proporción de varianza/correlación explicada por cada componente. Para realizar esto se divide lambda entre la varianza/correlación total (las lambdas están en eigen(S)[1]). La varianza/correlación total es la suma de las varianza/correlación de la diagonal de S. Una forma es sum(diag(S)).

La varianza/correlación total de los componentes es la suma de los valores propios (es decir, la suma de la varianza/correlación de cada componente). Las combinaciones lineales intentan replicar la varianza de X.

```

## [1] 0.7264164
## [1] 0.2036603
## [1] 0.04744993
## [1] 0.02216964
## [1] 0.0002104347
## [1] 8.143992e-05
## [1] 9.723436e-06
## [1] 1.497593e-06
## [1] 6.266923e-07

```

Valores de varianza explicada, que tanto peso tienen las variables al modelo.

Abajo se muestra el acumulado de los resultados anteriores sumando los primeros dos componentes

```
## [1] 0.7264164 0.9300767 0.9775266 0.9996963 0.9999067 0.9999882 0.9999979
## [8] 0.9999994 1.0000000

## [1] 0.5939898
## [1] 0.1356564
## [1] 0.1158368
## [1] 0.07420704
## [1] 0.03730141
## [1] 0.02321531
## [1] 0.01191711
## [1] 0.005781252
## [1] 0.002094814
```

Valores de correlación, que tanto peso tienen las variables al modelo. Abajo se muestra el acumulado de los resultados anteriores sumando los primeros dos componentes

```
## [1] 0.5939898 0.7296462 0.8454831 0.9196901 0.9569915 0.9802068 0.9921239
## [8] 0.9979052 1.0000000

## Warning: package 'FactoMineR' was built under R version 4.2.2

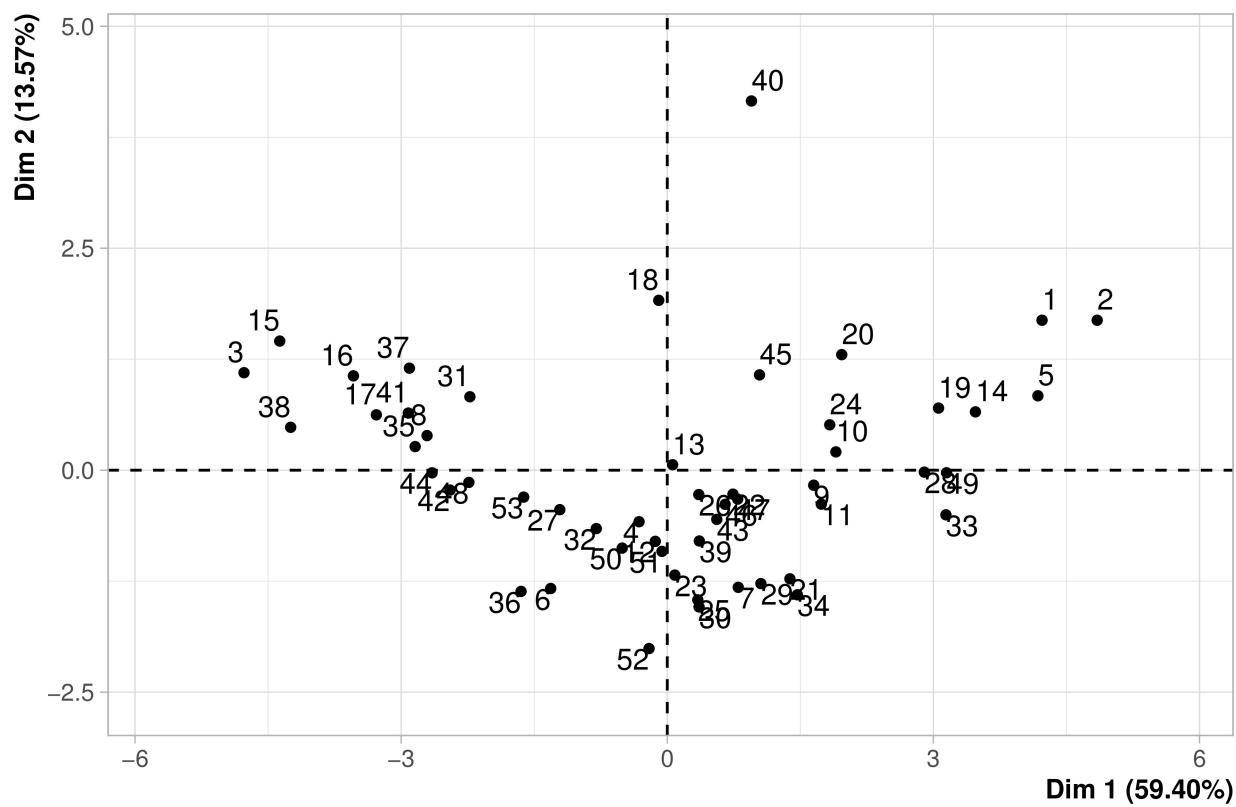
## Warning: package 'factoextra' was built under R version 4.2.2

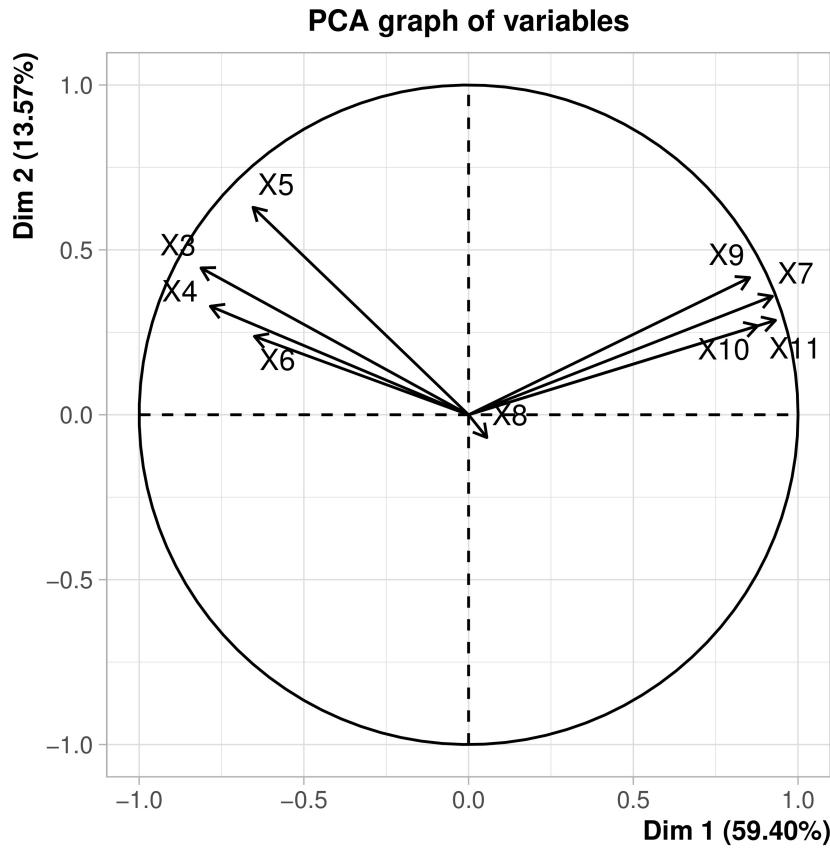
## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.2.2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

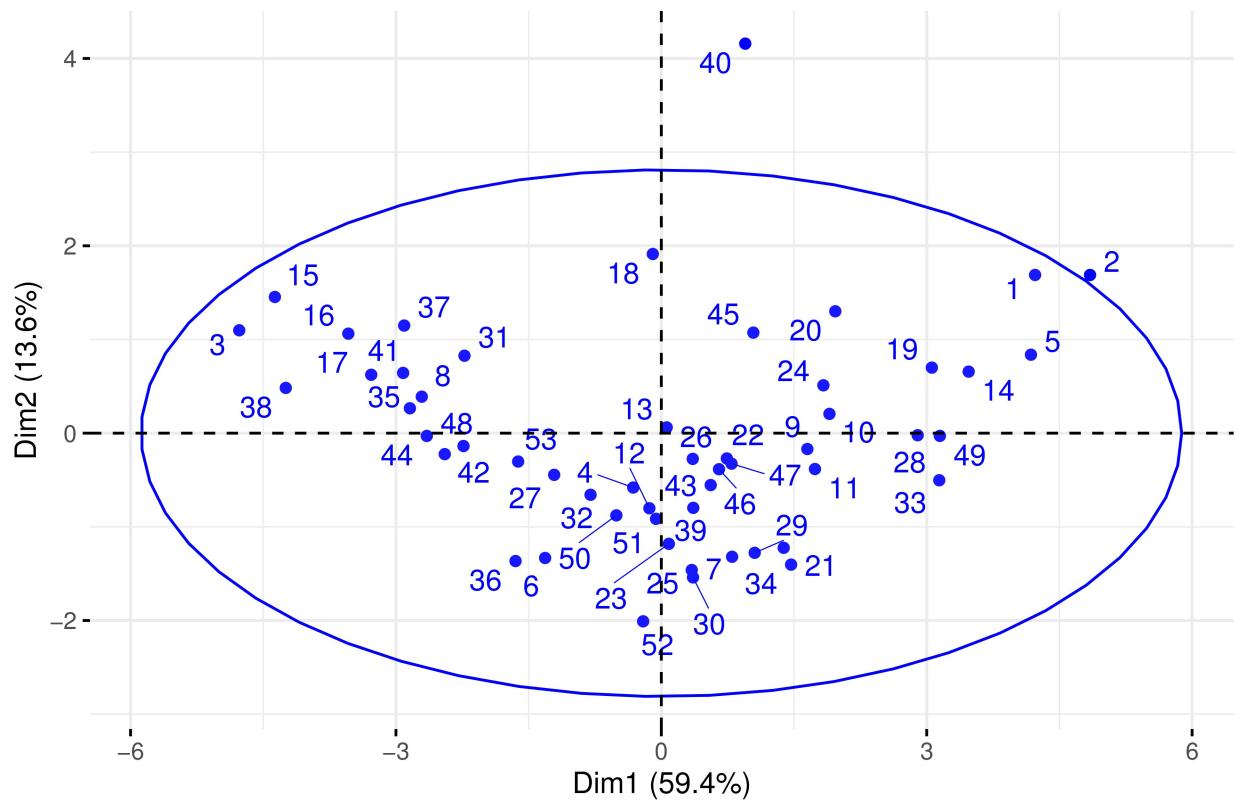
PCA graph of individuals





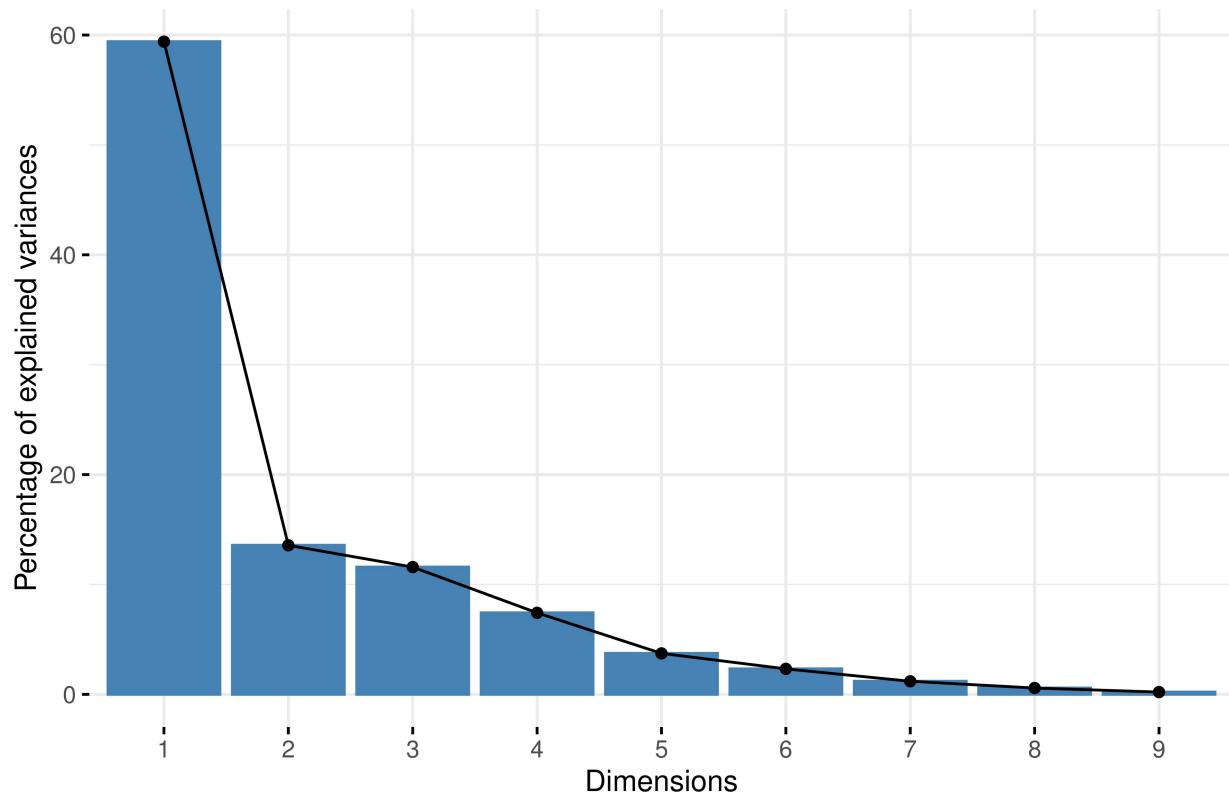
En el segundo gráfico muestra los vectores generados de cada componente respecto a dos dimensiones. Se puede apreciar que hay tres direcciones principales, en el primer cuadrante se observa el primer grupo integrado por variables que corresponden a factores que afectan el agua(X3, X4, X5 y X6), tales como alcalinidad, ph, calcio y clorofila. El segundo por factores relacionados a la concentración de mercurio en los peces(X7, X9, X10 y X11) y la tercera dirección va sobre la variable de la cantidad de peces estudiados. Por otro lado el tamaño del vector indica el aporte de cada componente al modelo con lo cual la dirección que indica el componente de la variable X8 no es muy relevante al modelo.

## Individuals – PCA



En el primer gráfico se observa la distancia de mahalanobis, los datos que se encuentras por afuera de la circunferencia de la elipse son los datos atípicos de la distancia de mahalanobis.

### Scree plot



En conclusión se utilizan dos modelos porque a partir del segundo componente se observa un cambio abrupto en la linea de tendencia que sigue la metrica de porcentaje de explicación de varianza. Debido a que la expliación del modelo estabiliza su incremento a partir del segundo componente. Otra razón importante es que el tercer componente se comprobó en la gráfica de vectores que no es muy significativa para el modelo. Utilizando un componente que agrupa variables de concentración de mercurio en los peces y otro que agrupa variables de la condición del agua.

[https://github.com/a01634928/TC3007C-501-A01634928/blob/main/E1\\_Portafolio\\_analisis/reto\\_pececitos\\_modulo2.Rmd](https://github.com/a01634928/TC3007C-501-A01634928/blob/main/E1_Portafolio_analisis/reto_pececitos_modulo2.Rmd)