

732A97 Multivariate Statistics Lab 1

Raymond Sseguya

2019-11-27

Question 1: Describing individual variables

Consider the data set in the T1-9.dat file, National track records for women. For 55 different countries we have the national records for 7 variables (100, 200, 400, 800, 1500, 3000m and marathon). Use R to do the following analyses.

input

```
trackrcs <- read.table("T1-9.dat",
  col.names = c("countries", "x100m", "x200m", "x400m", "x800m", "x1500m", "x3000m", "marathon"))

trackrcs2 <- (trackrcs)[,-1]
rownames(trackrcs2) <- trackrcs[,1]
```

a) Describe the 7 variables with mean values, standard deviations e.t.c

```
apply(trackrcs[, -1], 2, FUN = function(cl){
  c("mean"=mean(cl),
    "median"=median(cl),
    "sd"=sd(cl),
    "95% quantile"=quantile(cl, probs = 0.975, names = FALSE),
    "5% quantile"=quantile(cl, probs = 0.025, names = FALSE),
    "max"=max(cl),
    "min"=min(cl))
}) %>% t()
```

##	mean	median	sd	95% quantile	5% quantile	max
## x100m	11.357778	11.325	0.39410116	12.29875	10.74300	12.52
## x200m	23.118519	22.980	0.92902547	25.33625	21.76200	25.91
## x400m	51.989074	51.645	2.59720188	56.29075	48.07450	61.65
## x800m	2.022407	2.005	0.08687304	2.26700	1.91325	2.29
## x1500m	4.189444	4.100	0.27236502	4.75500	3.87975	5.42
## x3000m	9.080741	8.845	0.81532689	10.81075	8.36000	13.12
## marathon	153.619259	148.430	16.43989508	205.58625	138.76900	221.14
##	min					
## x100m	10.49					
## x200m	21.34					
## x400m	47.60					
## x800m	1.89					
## x1500m	3.84					
## x3000m	8.10					
## marathon	135.25					

b) Illustrate the variables with different graphs (explore what plotting possibilities R has). Make sure that the graphs look attractive (it is absolutely necessary to look at the labels, font sizes, point types). Are there any apparent extreme values? Do the variables seem normally distributed? Plot the best fitting (match the mean and standard deviation, i.e. method of moments) Gaussian density curve on the data's histogram. For the last part you may be interested in the `hist()` and `density()` functions.

```
j=0; lapply(trackrcs2, FUN = function(i){
  j <- j+1
  hist(i, main = colnames(trackrcs2)[j], xlab=colnames(trackrcs2)[j])
})
```

Question 2: Relationships between the variables

a) Compute the covariance and correlation matrices for the 7 variables. Is there any apparent structure in them? Save these matrices for future use.

```
C <- cov((trackrcs)[,-1])
corr_m <- cor((trackrcs)[,-1])
```

C

```
##           x100m      x200m      x400m      x800m      x1500m
## x100m      0.15531572  0.3445608  0.8912960  0.027703564  0.08389119
## x200m      0.34456080  0.8630883  2.1928363  0.066165898  0.20276331
## x400m      0.89129602  2.1928363  6.7454576  0.181807932  0.50917683
## x800m      0.02770356  0.0661659  0.1818079  0.007546925  0.02141457
## x1500m     0.08389119  0.2027633  0.5091768  0.021414570  0.07418270
## x3000m     0.23388281  0.5543502  1.4268158  0.061379315  0.21615514
## marathon  4.33417757 10.3849876 28.9037314 1.219654647 3.53983732
##           x3000m      marathon
## x100m      0.23388281  4.334178
## x200m      0.55435017 10.384988
## x400m      1.42681579 28.903731
## x800m      0.06137932  1.219655
## x1500m     0.21615514  3.539837
## x3000m     0.66475793 10.706091
## marathon 10.70609113 270.270150
```

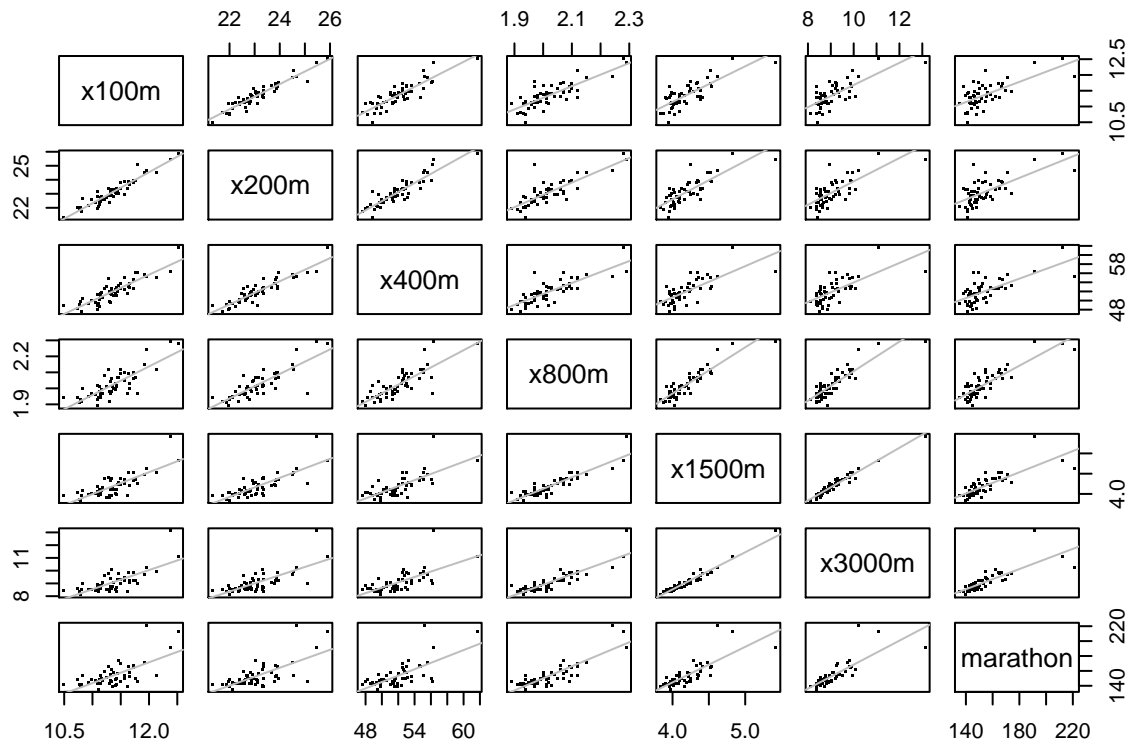
corr_m

```
##           x100m      x200m      x400m      x800m      x1500m      x3000m
## x100m      1.0000000  0.9410886  0.8707802  0.8091758  0.7815510  0.7278784
## x200m      0.9410886  1.0000000  0.9088096  0.8198258  0.8013282  0.7318546
## x400m      0.8707802  0.9088096  1.0000000  0.8057904  0.7197996  0.6737991
## x800m      0.8091758  0.8198258  0.8057904  1.0000000  0.9050509  0.8665732
## x1500m     0.7815510  0.8013282  0.7197996  0.9050509  1.0000000  0.9733801
## x3000m     0.7278784  0.7318546  0.6737991  0.8665732  0.9733801  1.0000000
## marathon  0.6689597  0.6799537  0.6769384  0.8539900  0.7905565  0.7987302
##           marathon
## x100m      0.6689597
## x200m      0.6799537
## x400m      0.6769384
## x800m      0.8539900
## x1500m     0.7905565
## x3000m     0.7987302
## marathon  1.0000000
```

Both matrices are symmetric. The correlation matrix has ones on the main diagonal.

b) Generate and study the scatterplots between each pair of variables. Any extreme values?

```
pairs(trackrcs[,-1], pch = ".", cex = 1.5, panel = function(x, y, ...){
  points(x, y, ...)
  abline(lm(y ~ x), col = "grey") })
```



The scatterplot matrix tells us that “marathon” is quite an outlier.

c) Explore what other plotting possibilities R offers for multivariate data. Present other (at least two) graphs that you find interesting with respect to this data set.

chernoff face

```
library(aplpack)
faces(trackrcs2)
```

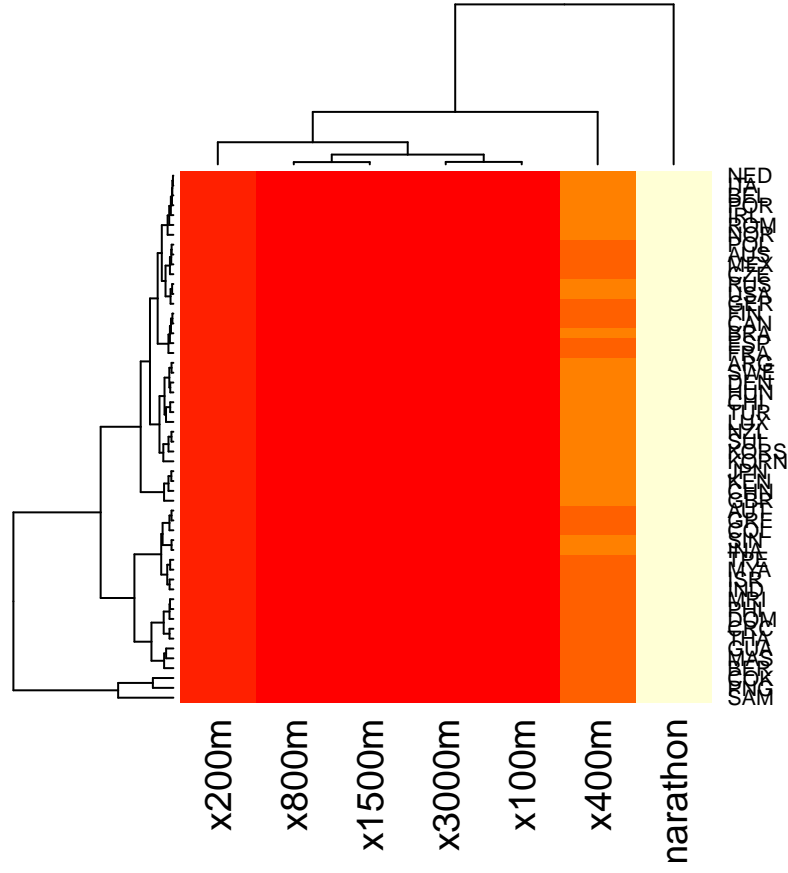
ARG	AUS	AUT	BEL	BER	BRA	CAN	CHI
CHN	COL	COK	CRC	CZE	DEN	DOM	FIN
FRA	GER	GBR	GRE	GUA	HUN	INA	IND
IRL	ISR	ITA	JPN	KEN	KORSKORN	LUX	
MAS	MRI	MEX	MYA	NED	NZL	NOR	PNG
PHI	POL	POR	ROM	RUS	SAM	SIN	ESP
SWE	SUI	TPE	THA	TUR	USA		

```
## effect of variables:
## modified item      Var
## "height of face   " "x100m"
## "width of face    " "x200m"
## "structure of face" "x400m"
## "height of mouth  " "x800m"
## "width of mouth   " "x1500m"
## "smiling          " "x3000m"
## "height of eyes   " "marathon"
## "width of eyes    " "x100m"
## "height of hair   " "x200m"
## "width of hair    " "x400m"
## "style of hair     " "x800m"
## "height of nose    " "x1500m"
## "width of nose     " "x3000m"
## "width of ear      " "marathon"
## "height of ear     " "x100m"
```

A stars plot is not very useful. We can tell from literature that it is also outdated.

heatmaps

```
heatmap(x=as.matrix(trackrcs2))
```



Question 3: Examining for extreme values

a) Look at the plots (esp. scatterplots) generated in the previous question. Which 3-4 countries appear most extreme? Why do you consider them extreme?

The 3 countries that seem most extreme are Brazil, Russia and the United States.

b) Compute the squared Euclidean distance (i.e. $r = 2$) of the observation from the sample mean for all 55 countries using R's matrix operations. First center the raw data by the means to get $(x - \bar{x})$ for each country. Then do a calculation with matrices that will result in a matrix that has on its diagonal the requested squared distance for each country. Copy this diagonal to a vector and report on the five most extreme countries. In this questions you MAY NOT use any loops.

```
x_bar = apply(trackrcs2,1,mean)
d0 = as.matrix(trackrcs2-x_bar); dim(d0)
```

```
## [1] 54 7
```

```
deviation = sqrt( d0%*%t(d0) ); dim(deviation)
```

```
## [1] 54 54
```

```
diagonal_vector <- diag(deviation)
deviation_countries <-
  cbind.data.frame(countries = as.vector(trackrcs[,1]),diagonal_vector)
deviation_countries_ordered <-
  deviation_countries[order(-deviation_countries$diagonal_vector), ]
deviation_countries_ordered[1:5,]
```

```
##      countries diagonal_vector
## PNG          PNG          193.0557
## COK          COK          185.0669
## SAM          SAM          165.7015
## BER          BER          151.3534
## GUA          GUA          148.7706
```

The five most extreme countries are PNG, COK, SAM, BER, GUA.

c)

```
V <- diag(apply(trackrcs2,2,var))
d_sq_v <- d0%%solve(V)%%t(d0)
diagonal_vector2 <- diag(d_sq_v)
deviation_countries2 <-
  cbind.data.frame(countries = as.vector(trackrcs[,1]),diagonal_vector2)
deviation_countries_ordered2 <-
  deviation_countries2[order(-deviation_countries2$diagonal_vector2), ]

deviation_countries_ordered2[1:5,]
```

```
##      countries diagonal_vector2
## COK      COK      302349.9
## PNG      PNG      302325.3
## SAM      SAM      256366.7
## GUA      GUA      213229.3
## BER      BER      211849.1
```

Except Great Britain, still the top five most extreme countries are the same and they are COK, PNG, SAM, GUA, BER

d) Compute the Mahalanobis distance, which countries are most extreme now?

```
d_sq_m <- d0%%solve(C)%%t(d0)
diagonal_vector3 <- diag(d_sq_m)
deviation_countries3 <-
  cbind.data.frame(countries = as.vector(trackrcs[,1]),diagonal_vector3)
deviation_countries_ordered3 <-
  deviation_countries3[order(-deviation_countries3$diagonal_vector3), ]

deviation_countries_ordered3[1:5,]
```

```
##      countries diagonal_vector
## COK      COK      185.0669
## PNG      PNG      193.0557
## SAM      SAM      165.7015
## GUA      GUA      148.7706
## BER      BER      151.3534
```

Still the top five most extreme countries are the same and they are COK, PNG, SAM, GUA, BER

e) Compare the results in b){d). Some of the countries are in the upper end with all the measures and perhaps they can be classified as extreme. Discuss this. But also notice the different measures give rather different results (how does Sweden behave?). Summarize this graphically. Produce Czekanowski's diagram using e.g. the RMaCzek package. In case of problems please describe them.

Sweden

Czekanowski's diagram

```
library(RMaCzek)
x<-czek_matrix(trackrcs2, n_classes = 7)
plot(x)
plot.czek_matrix(x)
```

Czekanowski's diagram

