

732A97 Multivariate Statistics Lab 1

Raymond Sseguya

2019-11-11

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")
colnames(trackrcs) <- c("countries", "100m", "200m", "400m", "800m", "1500m", "3000m", "marathon")
```

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

mean values

```
colMeans((trackrcs)[,-1])
```

```
##      100m      200m      400m      800m      1500m      3000m
## 11.357778 23.118519 51.989074  2.022407  4.189444  9.080741
##  marathon
## 153.619259
```

median

```
apply((trackrcs)[,-1], 2, median)
```

```
##      100m      200m      400m      800m      1500m      3000m  marathon
##   11.325   22.980   51.645    2.005    4.100    8.845  148.430
```

standard deviation

```
apply((trackrcs)[,-1], 2, sd)
```

```
##      100m      200m      400m      800m      1500m      3000m
## 0.39410116 0.92902547 2.59720188 0.08687304 0.27236502 0.81532689
##  marathon
## 16.43989508
```

maximum

```
apply((trackrcs)[,-1], 2, max)
```

```
##      100m      200m      400m      800m      1500m      3000m marathon
##      12.52      25.91      61.65      2.29      5.42      13.12      221.14
```

minimum

```
apply((trackrcs)[,-1], 2, min)
```

```
##      100m      200m      400m      800m      1500m      3000m marathon
##      10.49      21.34      47.60      1.89      3.84      8.10      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.

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.

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

cov_m
```

```
##           100m      200m      400m      800m      1500m
## 100m      0.15531572  0.3445608  0.8912960  0.027703564  0.08389119
## 200m      0.34456080  0.8630883  2.1928363  0.066165898  0.20276331
## 400m      0.89129602  2.1928363  6.7454576  0.181807932  0.50917683
## 800m      0.02770356  0.0661659  0.1818079  0.007546925  0.02141457
## 1500m     0.08389119  0.2027633  0.5091768  0.021414570  0.07418270
## 3000m     0.23388281  0.5543502  1.4268158  0.061379315  0.21615514
## marathon 4.33417757 10.3849876 28.9037314 1.219654647 3.53983732
##           3000m      marathon
## 100m      0.23388281  4.334178
## 200m      0.55435017 10.384988
## 400m      1.42681579 28.903731
## 800m      0.06137932  1.219655
## 1500m     0.21615514  3.539837
## 3000m     0.66475793 10.706091
## marathon 10.70609113 270.270150
```

```
corr_m
```

```
##           100m      200m      400m      800m      1500m      3000m
## 100m      1.0000000  0.9410886  0.8707802  0.8091758  0.7815510  0.7278784
## 200m      0.9410886  1.0000000  0.9088096  0.8198258  0.8013282  0.7318546
## 400m      0.8707802  0.9088096  1.0000000  0.8057904  0.7197996  0.6737991
## 800m      0.8091758  0.8198258  0.8057904  1.0000000  0.9050509  0.8665732
## 1500m     0.7815510  0.8013282  0.7197996  0.9050509  1.0000000  0.9733801
## 3000m     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
## 100m      0.6689597
## 200m      0.6799537
## 400m      0.6769384
## 800m      0.8539900
## 1500m     0.7905565
## 3000m     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)
```

