Exam 1 (Part 1)

David Lewis

March 25, 2025

Note that I used special rmarkdown syntax to embed some results directly into my prose. you can see the rmarkdown file for more details

Question 1

```r
data <- 90 * 1:100 - (1:100)^2 + 1000
```

What is the first, the seventeenth and the last entry of the vector data?

The first member of the vector is 1089. The 17th member of the vector is 2241. The last member of the vector is 0

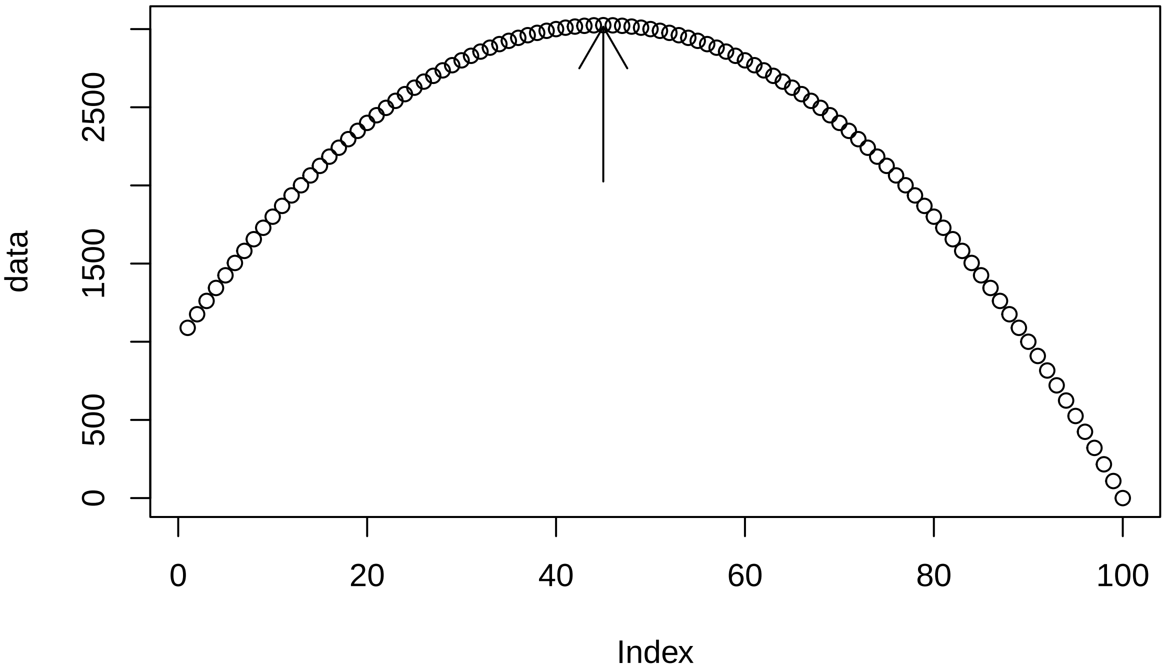
What is the maximum of the vector data? At which index is the maximum attained?

The maximum member of the vector is 3025. The index of this maximum is 45.

Plot the vector data with plot(data) and visually confirm your last result.

```r
plot(data)
y <- max(data)
x <- which.max(data)

arrows(x, y - 1000, x, y - 10)
```



The arrow points the max entry, confirming the last result.

At which indices are the entries of data between 2000 and 2500?

Assuming “between” is not inclusive.

```r
index <- which(data > 2000 & data < 2500)
```

The indices that are between 2000 and 2500 are 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77 .

Question 2

```r
m <- matrix(11:35, nrow = 5, byrow = TRUE)
```

What is the entry in the third row and forth column?

The entry in the third row and 4th column is 24

Briefly describe in words what `m[2:4,3:5]` returns.

The syntax `2:4` represents a range of numbers from 2 to 4, i.e (2,3,4). So in this case `m[2:4, 3:5]` is a submatrix of m containing the 2,3,4 rows and 3,4,5 columns.

```r
m[2:4, 3:5]
```

18	19	20
23	24	25
28	29	30

Question 3

Define the variable treatment as a vector of length 100 with elements: ("yes", "control", "yes", "control", . . . , "yes", "control")

```
treatment <- rep(c("yes", "control"), 50)
```

Define the variable smoker as a vector of length 100 with elements: ("yes", ”yes”, ”no”, ”no”, ”no”, . . . , "yes", "yes", "no", "no", "no")

```
smoker <- rep(c("yes", "yes", "no", "no", "no"), 20)
```

Define the vector :

```
lifespan <- abs(round(100 * sin(1:100)))
```

Create a data frame with treatment, smoker, and lifespan vectors. You may think of lifespan as the life span of the individuals.

```
data <- data.frame("treatment" = treatment, "smoker" = smoker, "lifespan" = lifespan)
```

Define a new vector x which consists of all elements of lifespan at whose index in smoker is the element "yes".

```
x <- data[data$smoker == "yes", ]$lifespan
```

What is the maximum of lifespan over all smokers?

The maximum lifespan over all smokers is 100

Half of the individuals got a certain treatment. Produce a new vector consisting of the lifespans of all individuals which are smokers and got the treatment.

```
lifespan_smokers_with_treatment <- data[data$smoker == "yes" & data$treatment == "yes", ]$lifespan
```

Question 4

using the Wages data from Ecdat library (also attached as Wages.csv), do the following: Show the top and bottom six rows of the data.

```
wages <- read.csv("Wages.csv")
wages$X <- NULL
```

```
rbind(head(n = 6, wages), tail(n = 6, wages))
```

	exp	wks	bluecol	ind	south	smsa	married	sex	union	ed	black	lwage
1	3	32	no	0	yes	no	yes	male	no	9	no	5.56068
2	4	43	no	0	yes	no	yes	male	no	9	no	5.72031

	exp	wks	bluecol	ind	south	smsa	married	sex	union	ed	black	lwage
3	5	40	no	0	yes	no	yes	male	no	9	no	5.99645
4	6	39	no	0	yes	no	yes	male	no	9	no	5.99645
5	7	42	no	1	yes	no	yes	male	no	9	no	6.06146
6	8	35	no	1	yes	no	yes	male	no	9	no	6.17379
4160	2	50	no	0	no	yes	no	female	no	12	no	5.85793
4161	3	50	no	0	no	yes	no	female	no	12	no	5.95324
4162	4	49	no	0	no	yes	no	female	no	12	no	6.06379
4163	5	50	no	0	no	yes	no	female	no	12	no	6.21461
4164	6	50	no	0	no	yes	no	female	no	12	no	6.29157
4165	7	50	no	0	no	yes	no	female	no	12	no	6.37161

How would you calculate the relative proportion (in percentages) of male and female workers?

```r
prop.table(table(wages$sex)) * 100
```

female	male
11.2605	88.7395

How do you calculate these proportions separately for workers in the south (yes, no)?

```r
prop.table(table(subset(wages, select = c("south", "sex"))), margin = 1) * 100
```

south/sex	female	male
no	10.21651	89.78349
yes	13.81307	86.18693

How do you calculate these proportions for all workers with an lwage larger than 6.5?

```r
x <- wages[wages$lwage > 6.5, ]

prop.table(table(subset(x, select = c("south", "sex"))), margin = 1) * 100
```

south/sex	female	male
no	4.820416	95.17958
yes	4.689864	95.31014

Question 5

You again use the dataset Wages. Use the command aggregate() to construct a data frame with three columns: In one column you have ed, in the second column you have sex, and in the third column you have for each possible combination of ed and sex the median of lwage for this combination. Since there are only 14 levels of ed and only two levels of sex, your new data frame will have at most 28 rows. How many rows does your data frame actually have? Why does it have fewer than 28 rows?

```
aggregate_lwage <- aggregate(lwage ~ ed + sex, data = wages, FUN = median)
```

The data frame only has 25 rows. This is because some combinations of the data do not exist. In this case (ed:4, sex:female), (ed:5, sex:female) and (ed:6, sex:female) do not exist. This is the reason why there are 3 less rows.