Homework 1

David Lewis

January 30, 2025

load the data

```
studentdata <- LearnBayes::studentdata
```

1. Size of the data

```
cat(paste("number of rows=", nrow(studentdata), "\nnumber of columns=", ncol(studentdata), sep=""))
```

```
## number of rows=657
## number of columns=11
```

2. Top 6 rows of the data

```
head(studentdata, n=6)
```

```
##   Student Height Gender Shoes Number Dvds ToSleep WakeUp Haircut  Job Drink
## 1       1     67 female    10      5   10    -2.5    5.5      60 30.0 water
## 2       2     64 female    20      7    5     1.5    8.0       0 20.0   pop
## 3       3     61 female    12      2    6    -1.5    7.5      48  0.0  milk
## 4       4     61 female     3      6   40     2.0    8.5      10  0.0 water
## 5       5     70   male     4      5    6     0.0    9.0      15 17.5   pop
## 6       6     63 female    NA      3    5     1.0    8.5      25  0.0 water
```

3. Pull out the complete data of Students 4, 22, 517, and 533.

```
studentdata[studentdata["Student"] %in% c(4, 22, 517, 533)]
```

```
## data frame with 0 columns and 657 rows
```

4. Using the documentation command, describe each variable in the data.

```
library(printr)
?LearnBayes::studentdata
```

```
## Student dataset
##
## Description:
##
##      Answers to a sheet of questions given to a large number of
##      students in introductory statistics classes
##
## Usage:
```

```
##
##      studentdata
##
## Format:
##
##      A data frame with 657 observations on the following 11 variables.
##
##      Student student number
##
##      Height height in inches
##
##      Gender gender
##
##      Shoes number of pairs of shoes owned
##
##      Number number chosen between 1 and 10
##
##      Dvds name of movie dvds owned
##
##      ToSleep time the person went to sleep the previous night (hours
##           past midnight)
##
##      WakeUp time the person woke up the next morning
##
##      Haircut cost of last haircut including tip
##
##      Job number of hours working on a job per week
##
##      Drink usual drink at suppertime among milk, water, and pop
##
## Source:
##
##      Collected by the author during the Fall 2006 semester.
```

The description for each variable is under the "format" section of the above output.

5. What is the nature of each variable?

```
data.frame("class" = sapply(studentdata, class))
```

	class
Student	integer
Height	numeric
Gender	factor
Shoes	numeric
Number	integer
Dvds	numeric
ToSleep	numeric
WakeUp	numeric
Haircut	numeric
Job	numeric
Drink	factor

6. Show the summary statistics of each variable.

```
summary(studentdata)
```

Student	Height	Gender	Shoes	Number	Dvds	ToSleep	WakeUp	Haircut	Job	Drink
Min. : 1	Min. :54.0	female:435	Min. : 0.00	Min. : 1.00	Min. : 0.00	Min. :-2.500	Min. : 1.000	Min. : 0.00	Min. : 0.00	milk :113
1st Qu.:165	1st Qu.:64.0	male :222	1st Qu.: 6.00	1st Qu.: 4.00	1st Qu.: 10.00	1st Qu.: 0.000	1st Qu.: 7.500	1st Qu.: 10.00	1st Qu.: 0.00	pop :178
Median :329	Median :66.0	NA	Median : 12.00	Median : 6.00	Median : 20.00	Median : 1.000	Median : 8.500	Median : 16.00	Median :10.50	water:355
Mean :329	Mean :66.7	NA	Mean : 15.42	Mean : 5.67	Mean : 30.93	Mean : 1.001	Mean : 8.383	Mean : 25.91	Mean :11.45	NA's : 11
3rd Qu.:493	3rd Qu.:70.0	NA	3rd Qu.: 20.00	3rd Qu.: 7.00	3rd Qu.: 30.00	3rd Qu.: 2.000	3rd Qu.: 9.000	3rd Qu.: 30.00	3rd Qu.:17.50	NA
Max. :657	Max. :84.0	NA	Max. :164.00	Max. :10.00	Max. :1000.00	Max. : 6.000	Max. :13.000	Max. :180.00	Max. :80.00	NA
NA	NA's :10	NA	NA's :22	NA's :2	NA's :16	NA's :3	NA's :2	NA's :20	NA's :32	NA

7. What is the gender distribution?

```
summary(studentdata$Gender)
```

```
## female   male
##    435    222
```

8. What is the most favorite number of the students?

```
sort(table(studentdata$Number), decreasing=TRUE)
```

7	5	3	8	6	4	2	9	1	10
191	78	76	69	66	60	57	42	9	7

The most favorite number of the students is 7

9. What is the second most favorite number of the students?

The second favorite number is 5.

10. What is the least favorite number of students?

The least favorite number is 10.

11. Use the 'table' command on 'studentdata$Shoes' and show the output.

```
table(studentdata$Shoes)
```

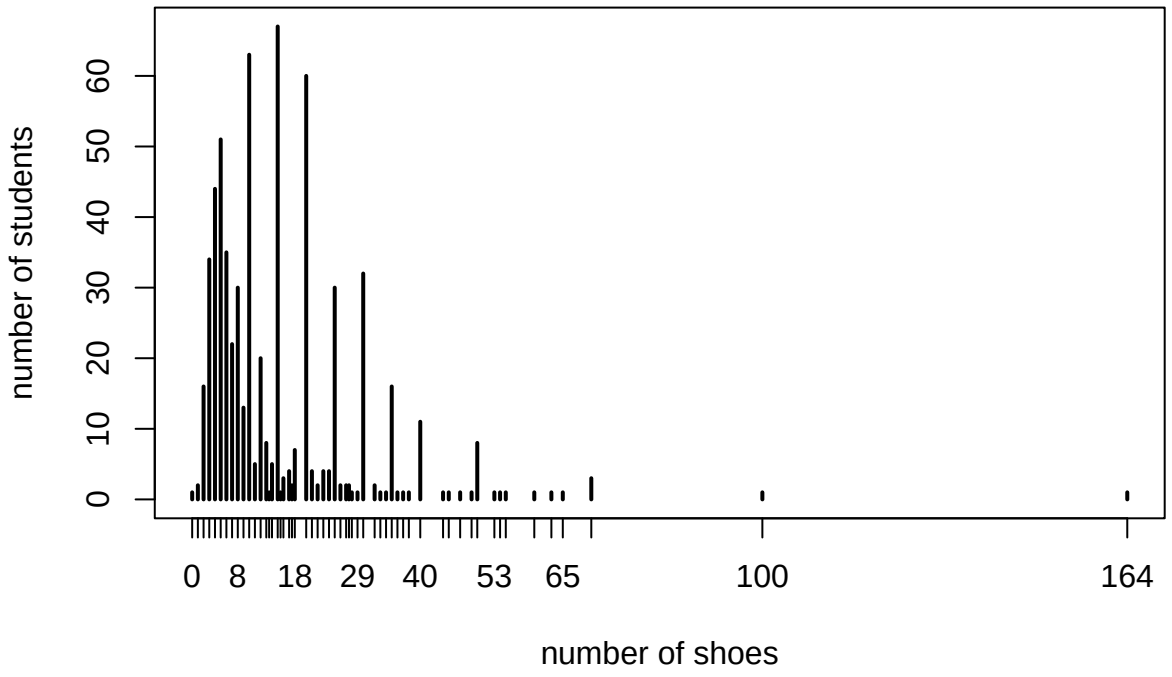
0	1	2	3	4	5	6	7	8	9	10	11	12	13	13.5	14	15	15.5	16	17	17.5	18	20	21	22	23	24	25	26	27	27.5	28	29	30	32	33	34	35	36	37	38	40	44	45	47
1	2	16	34	44	51	35	22	30	13	6	3	5	20	8	1	5	6	7	1	3	4	2	7	60	4	2	4	4	30	2	2	2	1	1	32	2	1	1	16	1	1	1	11	1	1	1

12. What is unusual about the output in 11?

One interesting thing is that some students seem to have an extra shoe that does not belong to a pair.

These data also appear to be spread out in a strange way. While it is true that the second most common number of pairs of shoes is 10, and most other students have a similar number of shoes, there are a relatively large number of students with 20 pairs of shoes.

```
plot(table(studentdata$Shoes), xlab="number of shoes", ylab="number of students")
```



Plotting the data shows that while much of the data fits a gaussian distribution, there are many "spikes" that don't fit the distribution.

13. Use the 'table' command on 'studentdata$Drink.'

```
table(studentdata$Drink)
```

milk	pop	water
113	178	355

14. Cross-tabulate 'Gender' and 'Drink.'

```
genderDrink <- table(studentdata$Gender, studentdata$Drink)
```

15. Calculate proportions row-wise and column-wise as well and correct to two decimal places.

Row-wise

```
round(prop.table(genderDrink, 1), 2)
```

/	milk	pop	water
female	0.15	0.26	0.60
male	0.23	0.31	0.46

Column-wise

```
round(prop.table(genderDrink, 2), 2)
```

/	milk	pop	water
female	0.56	0.62	0.72
male	0.44	0.38	0.28