Data Assignment 2

Yash Patel

25 February 2022

Question 1

You decide to test the impact of caffeine intake on dream vividness. You measure the average daily amount caffeine per day (in ounces) of a group. You also obtain (average) dream vividness ratings. The data looks like –

```
library(knitr)
df1 <- data.frame(
  Participant = 1:10,
  Caffeine.Intake = c(24, 16, 20, 28, 14, 10, 6, 20, 23, 28),
  Dream.Vividness.Rating = c(3, 7, 7, 7, 4, 4, 5, 7, 4, 6)
)
kable(df1)
```

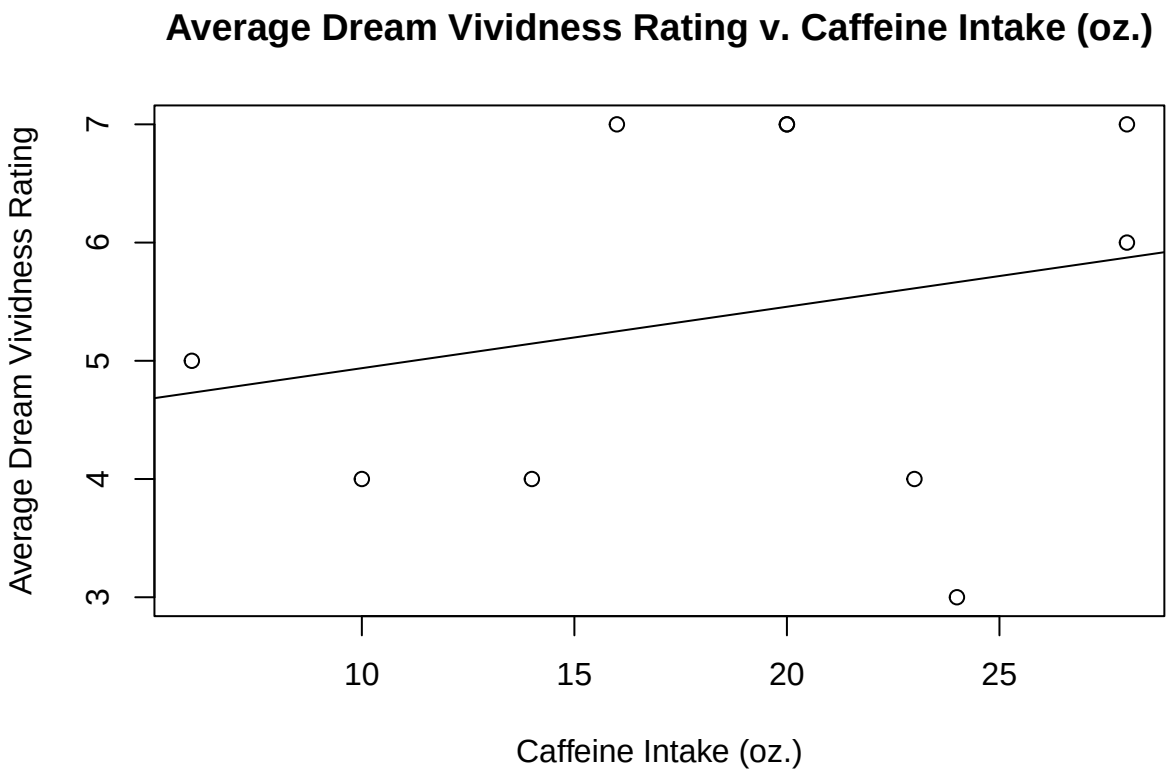
Participant	Caffeine.Intake	Dream.Vividness.Rating
1	24	3
2	16	7
3	20	7
4	28	7
5	14	4
6	10	4
7	6	5
8	20	7
9	23	4
10	28	6

What plot would you use to best express the data?

A simple scatterplot should be used with Caffeine.Intake on the x-axis and the Dream.Vividness.Rating on the y-axis.

Create a plot of caffeine intake against dream vividness rating

```
plot(df1$Caffeine.Intake, df1$Dream.Vividness.Rating,
     main = "Average Dream Vividness Rating v. Caffeine Intake (oz.)",
     xlab = "Caffeine Intake (oz.)",
     ylab = "Average Dream Vividness Rating")
abline(lm(df1$Dream.Vividness.Rating ~ df1$Caffeine.Intake))
```



Compute a correlation coefficient

```
cor(df1$Caffeine.Intake, df1$Dream.Vividness.Rating, method = "pearson")
```

```
## [1] 0.2427167
```

Interpret the data

With r = 0.2427167, there is a positive correlation between caffeine intake and average dream vividness rating, however this correlation is fairly weak.

Question 2

You run a personality inventory on 15 people and assess each person's level of "open-mindedness". The ratings range from 1 (not at all open-minded) to 100 (completely open minded). You get the following data –

```
library(knitr)
df2 <- data.frame(
  Participant = 1:10,
  Rating = c(94, 65, 46, 79, 69, 12, 67, 22, 47, 79)
)
kable(df2)
```

Participant	Rating
1	94
2	65
3	46
4	79
5	69
6	12
7	67
8	22
9	47
10	79

Compute the mean, median and mode for the data

```
mean(df2$Rating)
```

```
## [1] 58
```

```
median(df2$Rating)
```

```
## [1] 66
```

```
# Function obtained from StackOverflow since
# the default 'mode' function gives the
# internal representation type of objects
# rather than the statistical mode
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

getmode(df2$Rating)
```

```
## [1] 79
```

Compute the standard deviation (population and sample)

```
# This function takes the sample's standard deviation and calculates the population standard deviation,
sd.pop=function(x){
  sqrt((length(x)-1)/length(x)) * sd(x)
}

# Population SD:
# Note: This is assuming that the population mean is the same as the sample mean.
sd.pop(df2$Rating)
```

```
## [1] 24.71032
```

```
# Sample SD:
sd(df2$Rating)
```

```
## [1] 26.04697
```

Which standard deviation is the more appropriate one to use? Why?

The sample sd is more appropriate since this dataset is a subset of a larger group (population) that we want to make a conclusion about.

What r command would you use to compute the variance of the data?

```
var(df2$Rating)
```

```
## [1] 678.4444
```

Question 3

You are asked to analyze data from a study that looked at the association between an organizational self-rating ("how organized would you rate yourself?") and social network (how many "friends" would you say you currently have at this moment?") for 15 people. The data is –

```
library(knitr)
df3 <- data.frame(
  Participant = 1:15,
  Organizational.Self.Rating = c(9, 81, 57, 59, 51, 48, 32, 82, 44, 19, 10, 54, 20, 18, 15),
  Number.Of.Friends = c(14, 12, 6, 9, 5, 19, 16, 7, 17, 20, 15, 13, 11, 3, 4)
)
kable(df3)
```

Participant	Organizational.Self.Rating	Number.Of.Friends
1	9	14
2	81	12
3	57	6
4	59	9
5	51	5
6	48	19
7	32	16
8	82	7
9	44	17
10	19	20
11	10	15
12	54	13
13	20	11
14	18	3
15	15	4

Run a correlation on the data. What's Pearson's r?

```
cor(df3$Organizational.Self.Rating, df3$Number.Of.Friends, method = "pearson")
```

```
## [1] -0.1536362
```

Interpret the data

With r = -0.1536362, there is a extremely slight negative correlation between a person's organization self rating and the number of friends they have.

Relist the correlation by rank order below

```
library(knitr)
df3ranked <- df3
df3ranked$O.S.R.Rank <- c(1, 14, 12, 13, 10, 9, 7, 15, 8, 5, 2, 11, 6, 4, 3)
df3ranked$N.O.F.Rank <- c(10, 8, 4, 6, 3, 14, 12, 5, 13, 15, 11, 9, 7, 1, 2)
kable(df3ranked)
```

Participant	Organizational.Self.Rating	Number.Of.Friends	O.S.R.Rank	N.O.F.Rank
1	9	14	1	10
2	81	12	14	8
3	57	6	12	4
4	59	9	13	6
5	51	5	10	3
6	48	19	9	14
7	32	16	7	12
8	82	7	15	5
9	44	17	8	13
10	19	20	5	15
11	10	15	2	11
12	54	13	11	9
13	20	11	6	7
14	18	3	4	1
15	15	4	3	2

Looking at the data, what do you anticipate about a Spearman's correlation analysis?

The data does not seem to have any correlation in rank just by looking at it, so I'd assume the Spearman's rho to be low as well. By looking at it, I can't tell the direction, but it will probably be in the same direction as Pearson's r, meaning it will most likely be negative.

Run a Spearman's correlation. What is the output?

```
cor(df3$Organizational.Self.Rating, df3$Number.Of.Friends, method = "spearman")
```

```
## [1] -0.1607143
```

How does it compare to Pearson's r?

Pearson's r (-0.1536362) and Spearman's rho (-0.1607143) are both similar in magnitude and the same in sign, which is expected since it came from the same data set.

Create a scatter plot of the original data

```
plot(df3$Organizational.Self.Rating, df3$Number.Of.Friends,
  main = "Number of Friends v. Organization Self Rating",
  xlab = "Organizational Self Rating",
  ylab = "Number of Friends"
)
abline(lm(df3$Number.Of.Friends ~ df3$Organizational.Self.Rating))
```

Using lm, find the linear model for the best fit line. What is the value of the slope? What is the value of the intercept?

```
lm(df3$Number.Of.Friends ~ df3$Organizational.Self.Rating)
```

```
##
## Call:
## lm(formula = df3$Number.Of.Friends ~ df3$Organizational.Self.Rating)
##
## Coefficients:
##                     (Intercept)  df3$Organizational.Self.Rating
##                        12.78696                         -0.03473
```

The slope is -0.03473. The intercept is 12.78696.

Using abline, place trendline/regression line on your scatterplot

See above.