

# Practice Session 2

## Part 1 : Measures of central tendency & Measures of spread for quantitative data

### 1.1 Calculating the Sample Standard Deviation by Hand

Here is the formula for the sample mean:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

And here is the formula for the sample standard deviation:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Using the above data, perform the following calculations. Complete the table

for calculating the sample standard deviation.

Cost \$\$\$	b. Deviations ( $x_i - \bar{x}$ )	c. Deviations squared $(x_i - \bar{x})^2$
850		
900		
1400		
1200		
1050		
750		
1250		
1050		
565		
1000		
a. mean = _____		

d. Sum of squared deviations  $\sum_{i=1}^n (x_i - \bar{x})^2 =$  [ ]

e. Sum of squared deviations divided by n - 1:  $\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} =$  [ ]

$\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$

f. Take the square root to get s:  $=$  [ ]  $= s$

## 1.2 Five-Number Summary

Using the numbers from the previous exercise, do the following:

- Find the 5-number summary (minimum, Q1, median, Q3, maximum)
- Check your work using R functions

```
# your code here
```

## 1.3 Boxplots

Consider the `mtcars` data set. This data is built into R, so you can access it directly; no downloads required! First, create a histogram of the variable `mpg`. Then create a boxplot of `mpg`.

1. How do these two plots compare?
2. Create a boxplots of `mpg` per number of cylinders `cyl`.

```
# your code here
```

## 1.4 Quantitative data : histograms and outliers

Generate histograms for each of the following data sets. Use the `$` command to access the individual data sets. For each histogram, add the mean to the plot using `abline()`. Do you see any potential outliers? Also calculate the five-number summary for each using R.

```
set.seed(999)
s2_data = data.frame(
  dat1 = -rchisq(1000, df = 1),
  dat2 = rchisq(1000, df = 1),
  dat3 = runif(1000),
  dat4 = rnorm(1000),
  dat5 = sample(c(rnorm(1000, mean = 2), rnorm(1000, mean = 10)), size = 1000)
)
```

## 1.5 Percentiles

Compute the 25th, 50th, and 75th percentile for the 5 data sets in the `s2_data` data.frame. Which has the smallest median? Which has the largest?

```
## your code here
```

## 1.6 Normal Distribution and +/- 2 Standard Deviations

The normal distribution (also known as the “bell-curve”) occurs very frequently in mathematics, statistics, and the natural and social sciences. Which of the 5 data sets in the `s2_data` data.frame appears to be normally distributed?

Using this data set, find the mean and standard deviation, then calculate 1 standard deviations above, and 1 standard deviations below the mean. What percentiles do these values correspond to?

**Note:** approximately, there is 68% of data between 1 standard deviations above, and 1 standard deviations below the mean, in the normal distribution.

```
## your code here
```

## 1.7 Z-Scores

Read the following description on Z-scores, then answer the question below.

### 5.1 Standardizing with z-Scores

Expressing a distance from the mean in standard deviations *standardizes* the performances. To **standardize** a value, we subtract the mean and then divide this difference by the standard deviation:

$$z = \frac{y - \bar{y}}{s}$$

#### NOTATION ALERT

We always use the letter  $z$  to denote values that have been standardized with the mean and standard deviation.

The values are called **standardized values**, and are commonly denoted with the letter  $z$ . Usually we just call them  **$z$ -scores**.

$z$ -scores measure the distance of a value from the mean in standard deviations. A  $z$ -score of 2 says that a data value is two standard deviations above the mean. It doesn't matter whether the original variable was measured in fathoms, dollars, or carats; those units don't apply to  $z$ -scores. Data values below the mean have negative  $z$ -scores, so a  $z$ -score of  $-1.6$  means that the data value was 1.6 standard deviations below the mean. Of course, regardless of the direction, the farther a data value is from the mean, the more unusual it is, so a  $z$ -score of  $-1.3$  is more extraordinary than a  $z$ -score of  $1.2$ .

- 15. Temperatures** A town's January high temperatures average  $2^{\circ}\text{C}$  with a standard deviation of  $6^{\circ}$ , while in July the mean high temperature is  $24^{\circ}$  and the standard deviation is  $5^{\circ}$ . In which month is it more unusual to have a day with a high temperature of  $13^{\circ}$ ? Explain.

```
## your code here
```

## Part 2 : The Relationship Between Two Quantitative Variables/ Correlation and Regression

You might use the formula of the correlation between two quantitative variables:

$$r_{xy} = \frac{1}{(n-1)s_z s_y} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Remember that the fitted regression line is defined by the equation:

- $\hat{y} = a + bx$ , or
- $\text{Response} = a + b \cdot (\text{Explanatory})$
- Residuals = observed - predicted =  $y - \hat{y}$

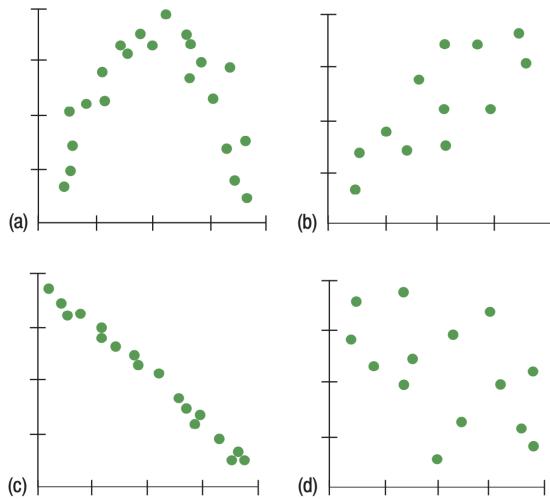
Where:

- Response: is the response variable or the dependent variable
- Explanatory: is the independent variable
- a: is the y-intercept
- b: is the slope of the regression line

You may use the following R functions: `plot()`, `lm()`, `cor()`, `abline()`. And you might need to download Lock5Data using `library(Lock5Data)`.

## 2.1 Describe scatterplots

Here are several scatterplots. The calculated correlations are 0.006, -0.977, -0.487, and 0.777. Match each scatter plot with the appropriate correlation coefficient.



**Answers:**

- a.
- b.
- c.
- d.

## 2.2 Create scatterplots in R

Load the data `FloridaLakes` from library(`Lock5Data`).

1. Describe the type of each of the variables: `pH`, `Calcium`, and `Alkalinity`.
2. Create a three scatter plots for each pair of variables: `pH vs Calcium`, `pH vs Alkalinity`, and `Calcium vs Alkalinity`. Add the main title to each plot.
3. What is the correlation coefficient between `pH` and `Calcium`? Is it positive or negative?
4. What do these coefficients mean in the context of this data?
5. Try to calculate the correlation coefficient between `pH` and `Calcium` without using the R function for correlation.

**Answers:**

```
# download the data and load it into R
library(Lock5Data)
data(FloridaLakes)
```

- 1.
- 2.
- 3.
- 4.
- 5.