STAT2005 Programming Languages for Statistics
Exercise for Chapter 2

1. Mary is living in an integer-valued one-dimensional space. She is initially at the origin at $x = 0$. She can either move to the left or right by 1 or 2 units in each second with equal probabilities. E.g. If Mary is at 0, she can move to -2, -1, 1, 2, in a second with equal probabilities.

(a) Using a random seed `2005`, write R codes to simulate Mary's movements during the first 120 seconds. Draw the path over time.

(b) Plot two red dashed horizontal lines at the maximum and minimum of the path.

2. (a) Using a random seed `2005`, generate 1,000 pseudorandom numbers from $X \sim N(3,4)$, store them in a vector named `x`.

(b) Generate 2,000 pseudorandom numbers from $Y \sim N(1,4)$, store them in a vector named `y`. Assume that the population mean and standard deviations of X and Y are unknown to us, but we know that their standard deviations are the same. Write R codes to find the pooled standard deviation of `x` and `y` and store it into a variable named `PooledSD`. The formula of pooled standard deviation is

$$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}},$$

where n_1, n_2 are the sample size of `x` and `y`, s_1^2 and s_2^2 are the sample variance of `x` and `y`.

(c) We are interested to perform a two-sample t-test for `x` and `y`.

$$H_0: \mu_1 = \mu_2 \text{ vs. } H_1: \mu_1 \neq \mu_2,$$

where μ_1, μ_2 are the unknown population means of X and Y. Write R codes to find the following t-statistics and compute the corresponding critical value at 95% significance level. The t-statistics is given by

$$t = \frac{(\bar{x}-\bar{y})}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2},$$

where t_{df} denotes the t-distribution with df degrees of freedom.

3. The file `ex2_q3.dat` stores the data of the GDP (Gross Domestic Product) for 26 countries (from "A" to "Z") in the first and second year. The variables are

`Country`: Country label from "A" to "Z";

`gdp1`: GDP of the first year;

`gdp2`: GDP of the second year;

`Region`: Each country belongs to one of the four regions "East", "South", "West" or "North".

(a) Write R codes to read `gdp.dat` as a `data.frame` object named `data`. Display the first 6 data.

(b) Write R codes to find the mean of `gdp1` in each `region`.

(c) Using the `by()` function, compute the sum of values in column `gdp1` and `gdp2`.

(d) Draw a scatter plot of `gdp1` and `gdp2` to show the relationship between GDP in the first and second year. Do you find any linear relation between the two?

1. (a).

```
set.seed(2025)
moves <- sample(c(-2,-1,1,2), size=200, replace=T, prob=c(0.25,0.25,0.25,0.25))
pos <- cumsum(c(0, moves))
pos <- as.ts(pos)
plot(pos)
```

(b)

```
max_pos <- max(pos)
min_pos <- min(pos)
abline(h=max_pos, col="red", lty="dashed")
abline(h=min_pos, col="red", lty="dashed")
```

2(a)

```
set.seed(2025)
x <- rnorm(1000, 3, 4)
```

(b)

```
y <- rnorm(2000, 1, 4)
n1 <- length(x)
n2 <- length(y)
s1_2 <- var(x)
s2_2 <- var(y)
Pooled SD <- sqrt((n1-1)*s1_2 + (n2-1)*s2_2) / (n1+n2-2))
```

(c)

```
x_mean <- mean(x)
y_mean <- mean(y)
t <- (x_mean - y_mean) / (Pooled SD * sqrt(1/n1 + 1/n2))
abs(t) > qt(0.975, n1+n2-2))
```

3. (a)

```
data <- read.table("gdp.dat", header=T)
head(data)
```

(b)

```
by(data$gdp1, data$region, mean)
```

for vector, use tapply Region //, use by is also OK

(c)

```
by(data[,c(2,3)], data$region, colSum)
rel).plot(data$gdp1, data$gdp2)
```