University of California, Los Angeles
Department of Statistics

Statistics 13 **Instructor: Nicolas Christou**

Homework 5

EXERCISE 1
Without performing all the calculations you should be able to indicate whether list P or list Q has the smaller sample standard deviation.

a. List P contains the 2000 integers
$1, 1, 2, 2, 3, 3, 4, 4, \cdots, 1000, 1000.$

List Q contains the 3000 integers
$1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, \cdots, 999, 999, 999, 1000, 1000, 1000.$

b. List P has 50 values.
The number 180 appears 10 times.

The number 200 appears 30 times.

The number 220 appears 10 times.

List Q has 50 values.
The number 180 appears 5 times.

The number 200 appears 40 times.

The number 220 appears 5 times.

c. List P contains 500 values.
The value 18 appears 100 times.

The value 19 appears 100 times.

The value 20 appears 100 times.

The value 21 appears 100 times.

The value 22 appears 100 times.

List Q contains 500 values.
The value 18 appears 150 times.

The value 20 appears 200 times.

The value 22 appears 150 times.

EXERCISE 2
In the the U.S. temperature is recorded in Farenheit degrees, while in most of the other countries it is recorded in Celcius degrees. Suppose a tourist from a country where temperature is recorded in Celcius degrees will visit Los Angeles this summer. He was told that the July average in Los Angeles is 85 Farenheit degrees, with a standard deviation of 10 Farenheit degrees. Help this tourist understand the weather conditions in Los Angeles.

EXERCISE 3
The sample variance is given by the following formula:

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

where n is the sample size and $\bar{x}$ is the sample mean. By expanding the above formula show that:

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

EXERCISE 4
You have just calculated that the sample mean and sample standard deviation of a sample of size 101 are 240.0 and 25.88 respectively. Unfortunately, a check of the list uncovers two errors:
A number listed as 230 should be 200.
A number listed as 250 should be 280.

a. After you make the corrections, what will be the sample mean?

b. What will be the standard deviation, after you make the correction?

EXERCISE 5

Use `R` to access the Maas river data. These data contain the concentration of lead and zinc in `ppm` at 155 locations at the banks of the Maas river in the Netherlands. You can read the data in `R` as follows:

```
soil <- read.table("http://www.stat.ucla.edu/~nchristo/statistics13/soil.txt", header=TRUE)
```

a. Compute the summary statistics for `lead` and `zinc`.

b. Plot the histogram of `lead` and `log(lead)`.

c. Plot `log(lead)` against `log(zinc)`. What do you observe?

d. The level of risk for surface soil based on lead concentration in `ppm` is given on the table below:

Mean concentration (`ppm`)	Level of risk
Below 150	Lead-free
Between 150-400	Lead-safe
Above 400	Significant environmental lead hazard

Use techniques similar to pages 9-10 in the handout on `R` to give different colors and sizes to the lead concentration at these 155 locations.

EXERCISE 6

The data for this exercise represent approximately the centers (given by longitude and latitude) of each one of the City of Los Angeles neighborhoods. See also the Los Angeles Times project on the City of Los Angeles neighborhoods at:

`http://projects.latimes.com/mapping-la/neighborhoods/`

You can access these data at:

```
a <- read.table("http://www.stat.ucla.edu/~nchristo/statistics13/
    la_data.txt", header=TRUE)
```

a. Plot these data points and add the map on the plot.

b. Do you see any relationship between income and school performance? Hint: Plot the variable `Schools` against the variable `Income` and describe what you see. Also, ignore the data points on the plot for which `Schools=0`.