Tutorial 1

Relevant material: Unit 1.

1. A group of 11 former college students are interviewed 10 years after their graduation. Their incomes are as follows (in 1 000 pounds):

$$\{20, 22, 23, 23, 25, 28, 28, 30, 30, 34, 160\}$$

For this sample, we have calculated the following summary statistics:

- Sample average ≈ 38.5
- Sample median 28
- Sample standard deviation ≈ 40.5
- Interquartile range 7 (from 23 to 30)

(a) We wish to measure the *central tendency* in this sample. Which measure is the most appropriate? (You can use the summary statistics provided above or other measures.) Argue.

(b) We wish to measure the *variability* in this sample. Which measure is the most appropriate? (You can use the summary statistics provided above or other measures.) Argue.

(c) Construct a box-and-whisker plot of the data.

(d) There was a reporting mistake in the data set - the largest value is actually 360 instead of 160. How do the summary statistics change? Do your answers to questions 1 and 2 change?

2. Consider the sample $\{X_1, X_2, \ldots, X_n\}$.

(a) Show that $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the solution of minimization problem

$$\min_c \sum_{i=1}^{n} (X_i - c)^2 , \tag{1}$$

(b) What is the interpretation of the function $\sum_{i=1}^{n} (X_i - c)^2$ in (1)?

(The aim of this exercise is justify the use of the sample average).

3. (a) Show that $\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2$.
What is this estimator?

(b) Show that $\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n) = \sum_{i=1}^{n}(X_i - \bar{X}_n)Y_i = \sum_{i=1}^{n} X_iY_i - n\bar{X}_n\bar{Y}_n$.
What is this estimator?