# Series 1

1. **Breakdown-Point:** Let the observations $x_1, \ldots, x_n$ be given, where $n = 2k+1$ is odd. Determine the breakdown points of

   **a)** the arithmetic mean $\left(\widehat{\mu} = \overline{x} = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} (x_i - \mu)^2\right)$

   **b)** the median $\left(\widehat{\mu} = \operatorname{med}(x_1, \ldots, x_n) = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} |x_i - \mu|\right)$.

2. **Confidence Intervals:** A new type of wheat was planted on nine plots of land. The harvest of these plots (in dz/ha) is given by

   35.6; 34.9; 36.0; 30.2; 36.2; 35.6; 35.8; 35.9; 36.1

   The data can be assumed to be a realization of 9 i.i.d. random variables. It can be read in with the following command:

   ```
   > scan(url("http://stat.ethz.ch/Teaching/Datasets/WBL/ertrag.dat"))
   ```

   **a)** Estimate the expected value of the data using the M-estimator $\hat{\mu}$ with Huber's $\psi$ function and $c = 1.345$ and compute the 95% confidence interval $\hat{\mu} \pm q^t_{0.975;n-1} se(\hat{\mu})$.
   **R-Hint:** You can use the function `huberM()` of the R-package `robustbase`.

   **b)** Now compute the classical confidence interval $\overline{x} \pm q^t_{0.975;n-1} se(\overline{x})$ for the expected value (using the arithmetic mean as estimator) and compare it to the robust confidence interval in **a)**.

3. **Different Linear Regressions:** We apply different regression methods on the data available at `http://stat.ethz.ch/Teaching/Datasets/WBL/oatsM16.dat`. Note that the explanatory variables `Block` and `Variety` are factor variables. The data has two response variables: the original variable `ValuesOrg` and a changed variable `Values` (5 values of the original variable have been replaced). You will need the package `MASS` for this exercise.

   **a)** **Linear Regression with Original Response:** Perform a linear regression analysis for the response variable `ValuesOrg` once using the classical OLS estimator and once using the robust Huber M-estimator, respectively. Compare the two fits in terms of the residual standard error, the normal plot and the $L_1$ distance between the estimated coefficients. Using the classical approach, are the two factor variables significant on the 5% level?
   **R-Hint:** You can perform regression with the Huber M-estimator with
   `rlm(formula, data = ..., psi = psi.huber, method = "M", maxit = 50)`

   **b)** **Linear Regression with Changed Response:** Perform the same analysis as in task **a)** but use the response variable `Values`. In addition, compare the estimated coefficients with the corresponding estimates in **a)**.

4. **Influence Function for a Simple Linear Regression:** The data available at `http://stat.ethz.ch/Teaching/Datasets/WBL/irisset.dat` contains petal length (variable `x`) and petal width (variable `y`) of Iris setosa plants. In the pairs-plot (length as x-axes and width as y-axis), we can see an extreme point, observation 42, with a width of 2.3. One might suspect that a mistake has been made. We would like to investigate the effect of such mistakes in a linear regression analysis:

   $$Y_i = \alpha + \beta \cdot x_i + E_i.$$

**a)** Make a pairs-plot and identify the outlier in the original data. Determine the estimator $\widehat{\beta}$ of the slope for four different values of $y_{42}$ and the corresponding values of the empirical influence function. How does the graph of the empirical influence function SC look like?

**Note:** Use the slightly modified definition of the empirical influence function:

$$SC(y_0;\, y_1, \ldots, y_n, T_n) = \frac{T_n(y_1, \ldots, y_{n-1}, y_0) - T_n(y_1, \ldots, y_n)}{1/n},$$

where $T_n$ is the estimator of $\widehat{\beta}$.
**R-Hint:** For $y_{42}$ we have set the values 2.5, 2.9, 3.3 and 4.1.

**b)** Determine the shape of the empirical influence function by investigating the dependency of the $\widehat{\beta}$-formula on $y_i$ (for instance on $y_{42}$).
As you know from regression analysis $\widehat{\beta}$ is estimated by

$$\widehat{\beta} = \frac{\sum_i (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

**c)** Draw the regression-lines for different $y_{42}$-values into the scatter-plot .
**R-Hint:**

```
plot(d.iris)
r.iris <- lm(y ~ x, data = d.iris); abline(r.iris)
##  y_42=2.5
d.iris[42, "y"] <- 2.5 ; t.lab <- 2
r.iris <- lm(y ~ x, data = d.iris)
abline(r.iris, lty = t.lab); points(d.iris[42, ], lty = t.lab, pch = t.lab)
```

**Exercise hour:** Monday, June 10, morning.