Предварительная обработка данных

В тех случаях, когда данные сильно различаются единицами измерения и/или диапазоном значений, перед применением того или иного метода анализа данных обычно применяют их преобразование:

1) **Нормализация данных** осуществляется по формуле:

$$x_i^{\text{нов.}} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

Очевидно, что в результате нормализации все данные принимают значения из диапазона [0, 1].

Для иллюстрации процедуры нормализации данных в R сгенерируем, например, выборку из нормального распределения с м.о. = 10 и средним. кв. отклонением = 3 и нормализуем данные:

```
mysample = rnorm(100,10,3)
```

```
 [1] 13.2200251  7.9409181  9.9012962  6.1313571  8.0451720  7.1364581
 [7]  9.1287330 13.3262100  8.0970256  9.9669826  6.9151867  6.1642959
[13]  8.0158325 13.7387233  8.2537628  8.2692797  7.2937547  7.2237249
[19]  8.6256698  5.6717384  5.9224207  7.4294040  3.7847371  5.8864629
[25]  5.3905950 13.9435247  3.9680841  8.4793425  9.3309028  6.6503167
[31]  7.1703631  8.6973843  9.7509685  7.1551783  6.0183940  1.2107638
[37]  6.2809253  8.4981879  9.2377078  9.8172324  7.3983250  8.2681441
[43] 10.3945229  9.4605796  4.7958279  9.0414654  8.2022818  2.2319289
[49] 11.4627419  7.3626964  9.4794264  9.7467132 12.6375183 17.5065891
[55] 11.3986092 14.3730055  4.7939419 10.8344460  9.5359158  7.1840826
```

```
xm
xM=max(mysample)
xM
range=xM-xm
range
```

```
> xm
[1] 3.564306
> xM=max(mysample)
> xM
[1] 16.91902
> range=xM-xm
> range
[1] 13.35471
```

```
mysample=(mysample-xm)/range
mysample
```

```
 [1] 0.23846822 0.57061750 0.24406194 0.16042715 0.6569634
 [7] 0.00347868 0.29745424 0.28622536 0.55764327 0.4965223
[13] 0.93846265 0.57741651 0.17341634 1.00000000 0.8186518
[19] 0.29342723 0.13652326 0.35439676 0.33207646 0.6376074
[25] 0.48134299 0.37475160 0.57481110 0.26016075 0.3109518
[31] 0.63815119 0.51449748 0.66970021 0.42315527 0.7811805
[37] 0.43798707 0.75131229 0.47848042 0.17124464 0.8892757
[43] 0.38375385 0.21830315 0.25429604 0.48366485 0.9762243
[49] 0.32443339 0.71979557 0.54053850 0.54706193 0.3817612
[55] 0.69258706 0.49083431 0.24917759 0.31758824 0.3976909
```

1) **Стандартизация (центрирование) данных** осуществляется по формуле:

$$x_i^{\text{нов.}} = \frac{x_i - \bar{x}}{s_x}. \qquad (2)$$