

Statistics in toxicology – Exercise sheet 1

Exercise 1: Descriptive analysis of the VPA dataset (15 points)

The valproic acid (VPA) dataset as described in the lecture contains gene expression measurements for 54675 genes. Seven positive concentrations 25, 150, 350, 450, 550, 800, 1000 μM were measured in three replicates, each, and the untreated control was measured in six replicates.

The R-dataset `VPADData-Random.Rda` contains the gene expression values for a sample of 500 genes. These are randomly chosen out of those 10000 genes with the highest variance across all 27 samples. The information about the measured concentration is given in the column names of the dataset. Conduct the following analyses (a) to (f) and interpret each result in 1-3 sentences.

- (a) Display the entire dataset via boxplots, analogously to the Figure from the lecture (Chapter 1.4, slide 22), i.e. with one boxplot per concentration and replicate.
- (b) For each gene, calculate the standard deviation of all three / six replicates corresponding to the same concentration. This yields 8 values for each gene. Visualize the results stratified by concentration, i.e. summarize the results via a histogram for each concentration separately.
- (c) For each concentration separately, find the gene for which the three / six replicates have the highest standard deviation. Plot the concentration-response profiles for the resulting eight genes. Repeat the analysis, now finding the gene for which the three / six replicates have the lowest standard deviation, for each concentration separately. Plot the concentration-response profiles for the resulting eight genes.
- (d) For each gene, calculate the mean expression value per concentration. This yields a reduced dataset with dimensions 500 and 8. Visualize the reduced dataset via boxplots analogously to (a), but now with only one plot per concentration.
- (e) Determine the number of genes for which the mean gene expression is monotonically increasing, and the number of genes for which it is monotonically decreasing. Plot the concentration-response profiles for all genes with monotonically increasing and monotonically decreasing profiles. Make sure to indicate both the individual measurements per concentration and the mean expression value per concentration in the plot.
- (f) For each gene, calculate the difference between the mean gene expression value of each positive concentration and the mean gene expression value of the control. Visualize the results with one histogram per concentration.

Upload the processed exercise sheets in the Moodle until Wednesday, 13 April 2021, 14:15. Please note the information listed in the Moodle on the submission formalities for this course.