Statistical Data Analysis in Excel: Practical Tasks

Please, take the data from http://edu.sablab.net/data/xls

Task 1. Frequency distribution [pancreatitis.xls]

The role of smoking in the etiology of pancreatitis has been recognized for many years. To provide estimates of the quantitative significance of these factors, a hospital-based study was carried out in eastern Massachusetts and Rhode Island between 1975 and 1979. 53 patients who had a hospital discharge diagnosis of pancreatitis were included in this unmatched case-control study. The control group consisted of 217 patients admitted for diseases other than those of the pancreas and biliary tract. Risk factor information was obtained from a standardized interview with each subject, conducted by a trained interviewer.

- Build the frequency distributions with respect to smoking of the patients in two groups: experimental and control

Use: *COUNTIF, SUM*

Task 2. Histogram, scatter plot, numerical measures, outliers [mice.xls]

The data set proposed here was downloaded from phenome database at http://phenome.jax.org/ and contains some phenotypical mouse parameters measured for 790 mice of 40 genetically diverse inbred strains. In the survey the blood chemistry, preference for calcium and sodium solutions in increasing concentrations, and bone and body composition were addressed. (Accession number: **MPD:103**)

- Build the histogram of the final weight of the mice (parameter "Ending weight")
- Build a scatter plot between starting and ending weights of the mice.
- Calculate mean, standard deviation, min, Q1, median, Q3, max for all the parameters
- Find the outliers in by "Bleeding time" parameters using z-score
- Calculate correlation between "Ending weight" and all other parameters.

Use: *AVERAGE, STD, MIN, MAX, MEDIAN, PERCENTILE, CORREL,*
Data Analysis → Histogram

Task 3. Probabilities and distributions

Assume that you have developed an extremely efficient chemical treatment for *glioblastoma*. During tests on animal models it was found that the substance X, which you use, is able to kill all tumor cells (theoretically), but being given at high concentration it leads to the death of a patient due to intoxication. As the survived cancer cells fast evolve into resistant form, the efficiency of the treatment is significantly reduced if the second course is given. Therefore the treatment should be performed in one injection.

The experimental data suggest that the average concentration needed for the positive treatment is **1** μg/kg. The concentration needed for effective treatment is, of course, a random variable. Being presented in log10 scale and in g/kg, it can be approximated by a normal random variable with mean of **–6** and standard deviation of **0.4**.

The 50% lethal dose for human is **35** μg/kg. And the tests on animals suggest that in log10 scale it has a normal distribution as well with the standard deviation of **0.3**.

- Plot the probability density functions for the lethal concentrations of X and a concentration needed for a success treatment of glioblastoma.
- Calculate the concentration of X which ensures that less then 0.1% lethal cases occurred due to treatment itself.
- Find the optimal concentration which maximizes the survival ratio among patients, assuming that the lethal dose and treatment efficiency are independent.

Use: LOG, NORMDIST,

Task 4. Interval estimation for proportion [pancreatitis.xls]

- Define a 95% confidence interval for never-smoking proportion of people coming to a hospital

Use: *NORMINV, SQRT*

Task 5. Interval estimation for mean [mice.xls]

- Define a 95% confidence intervals for the average ending weight of male and female mice from MA/MyJ strain.

Use: *TINV*

Task 6. Unpaired t-testing [mice.xls]

- Find the parameters which are significantly different for male and female populations

Use: *TTEST*

Task 7. Paired t-testing [bloodpressure.xls]

The systolic blood pressures of n=12 women between the ages of 20 and 35 were measured before and after usage of a newly developed oral contraceptive.

- Find the parameters which are significantly different for male and female populations

Use: *TTEST*

Task 8. ANOVA [depression.xls]

As part of a long-term study of individuals 65 years of age or older, sociologists and physicians at the Wentworth Medical Center in upstate New York investigated the relationship between geographic location and depression. A sample of 60 individuals, all in reasonably good health, was selected; 20 individuals were residents of Florida, 20 were residents of New York, and 20 were residents of North Carolina. Each of the individuals sampled was given a standardized test to measure depression. The data collected follow; higher test scores indicate higher levels of depression.

- Find the parameters which are significantly different for male and female populations

Use: Data Analysis → ANOVA: Single Factor, ANOVA: Two Factor With Replication

Task 9. Linear regression [cells.xls]

Assume that some cells are grown under different temperature conditions from 20º to 40º. A researched would like to find a dependency between T and cell number.

- Build linear regression between temperature and cell number. Define the 95% confidence intervals for the parameters

Use: Data Analysis → Regression

Task 10. Multiple comparison [all_data.xls]

Acute lymphoblastic leukemia (ALL), is a form of leukemia, or cancer of the white blood cells characterized by excess lymphoblasts. **all_data.xls** contains the results of full-trancript profiling for ALL patients and healthy donors using Affymetrix microarrays. The data were downloaded from ArrayExpress repository and normalized. The expression values in the table are in log2 scale.

- Define the group of genes differentially expressed in two conditions with FDR<0.05

Task 11. Confidence intervals for the random function

Two rates where measured for a PCR experiment: experimental value (X) and control (Y). 5 replicates where performed for each. From previous experience we know that the error between replicates is normally distributed.

- Provide an interval estimation for the fold change X/Y ($\alpha = 0.05$)
- Provide an interval estimation for the log fold change log2(X/Y)

Task 12. Goodness of Fit

The new treatment for a disease is tested on 200 patients. The outcomes are classified as: A – patient is completely treated; B – disease transforms into a chronic form; C – treatment is unsuccessful. In parallel the 100 patients treated with standard methods are observed.

- Check whether the distribution is changed due to the treatment