Course webpage: **http://edu.sablab.net/sdae2011**
Datasets: **http://edu.sablab.net/data/xls**

1. Descriptive Statistics. Probability Distributions

1.1. Work with ***mice*** data.
- Calculate the *mean*, *median* and *standard deviation* for the ending weight of female mice.
- Draw the histogram of bone mineral density.

1.2. The average weight of 60-day mice (*Mus musculus*) is 20g, with a standard deviation over population equal to 3g. You are ordering a mouse from animal facility for your experiment. Calculate the probability for this mouse to be lighter than 17g. Assume that weight distribution is normal.

2. Interval Estimation of Means and Proportions

2.1. Assume you have performed measurements of the lifetime of *Caenorhabditis elegans*. After observation of 9 nematodes you get an average lifetime of 15.4 days and standard deviation of 2.7 days. Calculate the standard deviation of the average lifetime (i.e. standard error).

2.2. To determine the frequency of type O blood (universal donor) in a population, a random sample of 100 people were blood typed for the ABO group. Of this 100, 42 were found to be type O. Using normal approximation, calculate 95% confidence interval for the proportion of the population that has type O blood.

2.3. ***cancer***. This data contains results of survey aimed at survival and life quality of patients with advanced lung cancer, performed by the North Central Cancer Treatment Group (Loprinzi CL et al, J. of Clinical Oncology. 12(3):601-7, 1994). Look for the survival time (column ***time***, given in days) and provide the 95% confidence interval for mean survival time independently for male and female patients. See *description* sheet in xls-file if necessary.

3. Testing Hypotheses about Means and Proportions

3.1. ***mice.*** Compare the mean weight changes for "C57BL/10J" and "A/J" strains. Provide proper hypotheses, p-value of the test and conclusion.

3.2. ***pdac***. This dataset contains the log expressions of 4 mRNAs measured over 209 samples of pancreatic ductal adenocarcinoma (cancer) and 44 samples of healthy pancreas. Compare the mean expressions of TGFBI in cancer and healthy samples.

3.3. ***tracemetal.*** Trace metals in drinking water affect the flavor of the water, and unusually high concentrations can pose a health hazard. Table shows trace-metal concentrations (zinc, in mg/L) for both surface water and bottom water at six different river locations. Does the concentration of Zn depend on the depth of water collection (bottom/surface)?

4. ANOVA

4.1. ***fertilizer***. In an agricultural experiment using barley, the effects of various commercial fertilizers and planting densities on yield were studied. Six different commercial fertilizers were used and the barley was planted in three different densities. The plots used were quite uniform with respect to soil characteristics, drainage, etc. The yields in kilograms per plot are recorded in the following table. Analyze theses data using ANOVA method.

- Which factor has higher effect on yield?
- Which combination of the fertilizer and planting density would you recommend to barley grower?

4.2. ***teeth.*** Dataset contains the result of an experiment, conducted to measure the effect of various doses of vitamin C on the tooth growth (model animal – Guinea pigs). Vitamin C and orange juice were given to animals in different quantities. Using 2-way ANOVA compare the effects due to treatment (vitamin, orange juice) and concentration. Provide a proper ANOVA model.

5. Linear Regression

5.1. ***pdac***. This dataset contains the log expressions of 4 mRNAs measured over 209 samples of pancreatic ductal adenocarcinoma (cancer) and 44 samples of healthy pancreas. Separately for each group (cancer, healty) build a linear regression between genes SERPINI2 and ZD52F10. Validate its statistical significance. For significant linear regression model – give 95% confidence interval for the slope.

5.2. ***leukemia.*** Patients were classified into the two groups according to the presence or absence of a morphologic characteristic of white cells. Patients termed AG positive were identified by the presence of Auer rods and/or significant granulature of the leukemic cells in the bone marrow at diagnosis. For AG-negative patients, these factors were absent. Leukemia is a cancer characterized by an overproliferation of white blood cells; the higher the white blood count (WBC), the more severe the disease. Separately for each morphologic group, AG positive and AG negative:

- Draw a scatter diagram to show a possible association between the log survival time (take the log yourself and use as the dependent variable) and the log WBC (take the log yourself) and check if a linear model is justified.
- Estimate the regression parameters, the survival time for a patient with a WBC of 20,000 (are estimates for different groups different?), and draw the regression line on the same graph with the scatter diagram.
- Calculate the coefficients of determination and provide your interpretation. Is there evidence of an effect modification for 2 groups?