# Assignment 1

The file **datacancer.xls** contains values for breast cancer mortality from 1950 to 1960 ($y$) and the adult white female population in 1960 ($x$) for 301 counties in North Carolina, South Carolina, and Georgia.
*(Source: Rice, 2007. Mathematical Statistics and Data Analysis.)*

1. Make a histogram of the population values for cancer mortality.
2. What are the mean of the population and of total cancer mortality? What are the variance and standard deviation of population?
3. Draw a scatter plot of the number of cases ($y$) versus population ($x$).
4. Compute the linear regression: $y = \beta_0 + \beta_1 x + \varepsilon$.
5. Do all necessary plots in order to investigate the necessary assumptions. What are your conclusions?
6. Plot the residuals versus log(population). Do you believe that the assumption of homoscedasticity is given in this problem?
7. Compute the following model: $\sqrt{deaths} = \beta_0 + \beta_1 \sqrt{inhabitants} + \varepsilon$. Hint: In SPSS use the tap 'Transformieren (Transform)' $->$ 'Variable berechnen (Compute variables)' to compute a new variable defined as deaths$^{0.5}$.
8. Investigate again the assumption of homoscedasticity. What are your conclusions?
9. Plot a QQ-Plot of the residuals as well as a histogram. What are your conclusions regarding the normality of the data?
10. Interpret the quality of the model.
11. Are the coefficients significant? How high is the impact of population on cancer mortality?

Put the obtained plots as well as your conclusions in a document together. *Send this file via email to me not later than* ***Wednesday, 30th March 2011, 12:00!***
This assignment is worth 15%.