Stat 359 Assignment 3

1. In a study examining smoking and lung cancer, a random sample of men between the ages of 55 and 60 was obtained. The smoking and disease status of each sampled subject was ascertained. For each subject, a '1' is assigned if the subject had lung cancer (case) and a '0' if not. Similarly, a '1' indicates that a subject is a smoker and a '0' indicates a nonsmoker. The data are found in the Excel file 'LungCancer'.

   - Read the data into R, and use table() function to produce a contingency table summarizing these data.
   - Assuming that there is no association between smoking and lung cancer, compute a table of 'expected' counts.
   - By hand, compute the observed value of the test statistic for testing association between lung cancer and smoking.
   - Assuming there is no association, what is the distribution of the test statistic?
   - Using R, compute the p-value for a test of association, and give a *detailed* conclusion based on the p-value and a comparison of the tables observed and expected counts.

2. The following data are from a study examining the incidence of tuberculosis in relation to blood groups in a sample of Eskimos. It is of interest to determine if there is any association between the disease and blood group within the ABO system.

Severity	**O**	**A**	**AB**	**B**
Moderate-advanced	7	7	7	13
Minimal	27	34	12	18
Not Present	55	52	11	24

   - Assuming that there is no association between disease and blood group, compute a table of 'expected' counts.
   - By hand, compute the observed value of the test statistic for testing association between disease and blood group.
   - Assuming there is no association, what is the distribution of the test statistic?
   - Using R, compute the p-value for a test of association, and give a *detailed* conclusion based on the p-value and a comparison of the tables observed and expected counts.

3. The file 'Anscombe' contains 4 different datasets, each of which are based on a response Y, and a covariate X.

   (a) Produce 4 scatter plots (one for each dataset), on the same page, illustrating the relationship between Y and X. Describe each of these briefly, and state if you think a linear model of the form $y_i = a + bx_i + \epsilon_i$ would be appropriate.

(b) Perform 4 separate simple linear regressions (one for each dataset)and produce a table (in your text editor (ie. word)) that shows the R^2 value. Discuss what is happening here (hint: for simple linear regression, R^2 is just the square of the sample correlation coefficient).

4. The file 'growth' gives data on the height of a white spruce tree measured annually for 50 years. Letting Y_t denote the height of the tree at year $t > 0$, we consider describing the growth of the tree over time with a non-linear model $Y_t = f(t) + \epsilon_t,\ \epsilon_t \overset{iid}{\sim} N(0, \sigma^2)$. Three growth curves are considered for $f(t)$

   (a) **Logistic:** $f(t) = a/(1 + b * \exp\{-ct\})$

   (b) **Gompertz:** $f(t) = a \exp\{-b \exp\{-ct\}\}$

   (c) **Von Bertalanffy:** $f(t) = a - a \exp\{-b(t + c)\}$

   - Fit all three models using the non-linear least squares function nls() in R. Explain how you are choosing the starting values for nls() in each case. Produce a figure depicting the estimated curves all on the *same* plot, along with the observed data. Be sure to include a legend to distinguish the different curves.
   - For each of the three models, give a 95% confidence interval for $lim_{t\to\infty} f(t)$. What does this represent?
   - Select the best of the three models, and plot an estimate of the derivative $\frac{df(t)}{dt}$, which represents the rate of growth over time.