Ma 322: Biostatistics
Data Analysis Projects

Prof. Wickerhauser

Due Monday, May 3rd, 2021

Choose one of the following projects. Prepare a one-page outline of your proposed data acquisition and analysis methods for the instructor's approval by **Monday, April 12th, 2021**. The outline should include:

- the project chosen,
- the hypotheses to test or the quantities to estimate,
- the data to be acquired, and
- the statistical methods to be employed.

Project reports should be *at most 10 pages long* and must be uploaded to CrowdMark by the due date. They should contain sections corresponding to the parts of the previously submitted outline, describing the data sources and methods in a manner sufficient to reproduce your results.

You may also design a project of your own choosing, subject to instructor approval, with the same outline, size, and completion deadlines. Please see the instructor during office hours, or by appointment, as soon as possible if you wish to take this option.

1. Cancer and ABO type: Bayesian analysis and correlations. The article at this URL contains data collected to test the hypothesis that ABO blood type is correlated with cancer type:

`http://ispub.com/IJPA/13/1/5982`

Additional information on ABO blood types and cancer may be found at

`https://en.wikipedia.org/wiki/ABO_blood_group_system`

`https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0184295`

Obtain the data on the blood types of cancer patients from the article and use it to test the independence of ABO phenotypes and cancer, or ABO phenotype and cancer type, or ABO allele counts on cancer or cancer type. Note: there are typographical errors in the contingency table of blood type counts by population, which you should detect by checking against the totals (which are accurate) and correct before performing your tests of independence.
To estimate allele counts from ABO phenotype frequencies, obtain and use WinBUGS, following the example starting on page 210 in Chapter 12 of our text, "MCMC Using BRugs." Allele proportions should be multiplied by the number of subjects and then by 2 (since the subjects are diploid) to get allele counts by population.
Tutorials on WinBUGS may be found at

`http://homepage.stat.uiowa.edu/~gwoodwor/BBIText/AppendixBWinbugs.pdf`

`http://www2.stat-athens.aueb.gr/~jbn/papers2/23b_Lykou_Ntzoufras_2011_Wires_WinBUGS_final.pdf`

You may use plots of the Dirichlet posterior pdfs generated by `ldhw()` to check the MCMC allele frequency estimates:

`http://www.math.wustl.edu/~victor/classes/ma322/r-eg-35.txt`

However, you must use MCMC to get full credit for this project.

2. Gene expression in cancer: Classification and regression trees. Many thousands of genes alter their expression in various cancers. These changes have the potential to identify early-stage cancers and also to aid in understanding carcinogenesis.
Microarray datasets mentioned in our text, Chapter 18, p.308 (NCI and SRBCT), are at these URLs:

`http://statweb.stanford.edu/~tibs/ElemStatLearn/data.html`

In addition, information on gene expression (gene chip) data may be found at these URLs:

`http://genome-www.stanford.edu/nci60/`

`http://genome-www.stanford.edu/nci60/help.shtml`

Using these gene expression datasets, build at least two classifiers for cancer type. Select a small number of genes using criteria other than those in the text example, such as maximal variance. Test your classifiers with cross-validation and give reasons why one classifier is preferable to another.