Cornell University
Cornell Statistical Consulting Unit

Stats Happening - March 2012

---

1. Three new statistical genetics related workshops in April.

Introduction to Statistical Genetics. This workshop aims to provide a bridge connecting genetics and statistics and will focus on the most fundamental problem of inferring phenotype-genotype relationships, a problem which is now encountered in many different fields of research. We will introduce basic biological terms and concepts, describe common types of data and questions researchers are confronted with, and the statistical methods used to address these types of questions. This workshop is being offered by Li Ma, PhD, on Wednesday April 4th from 3:00-4:30 pm in 100 Savage Hall. This workshop has **no pre-requisites** and is **free** to the entire Cornell community. Please pre-register and mark it in your calendars now! For more details and to register for this and the other workshops offered this semester go to: http://www.cscu.cornell.edu/workshops/schedule.php

Introduction to Microarray Data Analysis using Genespring. The purpose of this workshop is to introduce the basic process of microarray data analysis using Genespring GX, which is an industry platform for gene expression analysis. In this workshop you will have access to Genespring GX and gain knowledge through hands-on exercises, which will include importing and organizing data sets, setting up experiments such as specifying parameters and normalizing data, performing quality control, and running statistical tests, etc. This workshop will be offered Wednesday, April 18, 2012 from 2:00 PM - 4:00 PM in the Stone Computer Room in Mann Library. For more information or to register, please visit our web site at http://www.cscu.cornell.edu/workshops/registration.php

Analyzing Gene Expression Data in R. This hands-on workshop is a follow-up to our "Introduction to Statistical Genetics" and "Getting Started with Microarray Data Analysis using Genespring". We will demonstrate an alternative approach to processing gene expression data. We will introduce a free R package, explain its approach to finding differentially expressed genes, and show in detail how it can be used to process gene expression data, and how it controls the false discovery rate. **Knowing R is not a prerequisite for this workshop.** We will show, step by step, how to read the data, produce diagnostics plots, normalize and filter the data, perform the statistical analysis, and interpret the results. This workshop will be offered on Thursday, April 19, 2012 from 2:30 PM - 4:00 PM in the B30B computer lab in Mann Library. This workshop will be taught by the author of this R package, Haim Bar, PhD. For more information or to register, please visit our web site at:
http://www.cscu.cornell.edu/workshops/registration.php

2. Useful tools for Response Surface Methodology.

Many fields of study use designed experiments to determine an optimal response in relation to multiple predictor variables. One common approximation used to solve this optimization problem using second-degree polynomials is response surface methodology (RSM). Challenges occur in choosing the ideal experimental design and the proper data analysis. The statistical software package JMP offers user-friendly modules to address both. From the DOE drop-down menu select Response Surface Design and follow the menus to design the experiment. For the analysis choose the Fit Model command in the Analyze menu, enter your variables as usual, and choose Response Surface Effect within the Attributes option to designate the main effects as RS in the model. Explore the response surface in the JMP output using the Contour Profiler in the Factor Profiling menu.

For more detailed information about RSM in JMP, download the pdf from the authors at:
http://www2.sas.com/proceedings/sugi22/STATS/PAPER265.PDF

There is also an R package for RSM analysis, called ‘rsm’. The following paper will be very useful to R users interested in using the package:
http://www.jstatsoft.org/v32/i07/paper

SAS code for RSM has also been developed and posted on their website, although with minimal functionality in exploring the response surface:
http://support.sas.com/documentation/cdl/en/imlug/59656/HTML/default/viewer.htm#genstatexpls_sect9.htm

3. Caution when using log transformations in JMP.

It is very common practice for researchers to normalize variables via log transformations, and this can be easily achieved within any statistical software package. Within JMP there are two easy ways to log transform a variable to be used in a model: 1) create a new permanent variable that is the log of another permanent variable, which can be subsequently entered into a statistical model; and 2) use the Transform option within the Analyze/Fit Model menu to include a temporarily log transformed variable within the current statistical analysis. Ostensibly these methods should be identical. However there are unfortunate differences in results for the least square (LS) means when it is the outcome variable that is being log transformed. Specifically, when using the Analyze/Fit Model menu to log transform the outcome temporarily, the LS means are on the original (not log) scale, but the standard errors around those estimates are in log scale. JMP does not provide the correct standard errors around these back-transformed estimates. The table has a footnote denoting this mismatch but the footnote can easily be overlooked with the unfortunate results of reporting standard errors that are way too small. In contrast, if you have created a permanent log transformed outcome variable, the LS means will be on log scale as will the standard errors that JMP reports. To avoid any confusion we would suggest creating a new log transformed variable rather than use the available transformations in the fit model menu.

4. Free user-friendly statistical software: PSPP.

PSPP is a useful tool for people needing to perform basic statistical analyses with their data. Its appearance is very similar to SPSS, and importantly, it is **free**. PSPP works on most operating systems such as Windows, Linux and Mac OS X. PSPP has a drop-down menu interface but can also be used with a command-syntax interface. PSPP can read SPSS syntax files and import several types of data files, such as Excel spreadsheets, and SPSS data files, etc. It manipulates and analyzes data, and writes the results to a listing file or to a standard text output. PSPP procedures provide a basic set of analyses: descriptive statistics, t-tests, ANOVA, linear regression, non-parametric tests, factor analysis, etc. The syntax accepted by PSPP and SPSS are also similar. Future versions of PSPP are slated to provide a greater variety of statistical analyses.

For downloading PSPP and more detailed information, please go to:
http://www.gnu.org/software/pspp/pspp.html

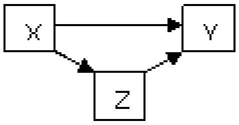
5. Assessing logistic regression model performance in R.

The most common method for analyzing data involving binary outcome variables is logistic regression. Assessment of the performance of such models can be done via dozens of related statistics, and now R has a package called ROCR that does nearly all. The package name derives from one of the most common techniques for assessing logistic models: the area under the Receiver Operator Characteristic (ROC) curve (see http://www.cscu.cornell.edu/news/statnews/stnews07.pdf for more details). ROCR utilizes the impressive graphical abilities of R to display ROC curves and many other graphical summaries, but also includes many helpful numeric summaries as well (e.g. Mathew's Correlation Coefficient). This package is easy to use (only three new R commands!), flexible, powerful, and offers impressive graphical representation.

Tobias Sing, Oliver Sander, Niko Beerenwinkel, Thomas Lengauer. ROCR: visualizing classifier performance in R. *Bioinformatics* 21(20):3940-3941 (2005).

6. Sample size and power calculation for testing mediation.

Mediation analysis is commonly used in the social sciences but has also proven to be a useful technique in other fields. Mediation occurs when an independent variable X affects a dependent variable Y partly or completely through an intermediate variable Z.



Calculating the sample size needed for such an analysis has become easier with the publication of the following article: Vittinghoff, Sen, & McCulloch. Sample size calculations for evaluating mediation.

Statistics In Medicine. 2009, 28: 541-557. The approach used is not only applicable to linear regression models but provides approximations for the Logistic, Poisson and Cox models. In addition to obtaining sample sizes, the methods presented by the authors can also be used to compute power or minimal detectable effects for a given sample size. The methods presented in the paper have been implemented in the new R package “powerMediation”, found at:
http://cran.r-project.org/web/packages/powerMediation/powerMediation.pdf .

A more limited SAS macro has also been developed, but it will only calculate the sample size needed for testing mediation in linear regression based on this paper. See:
http://analytics.ncsu.edu/sesug/2010/PO19.Kadel.pdf