Paper Assignment

(Due Thursday, December 22, 2005 @ 9am, electronically in PDF format or as a hard copy in Ingo's mail box).

Write up the results of your analysis of the proteomics data described below. Your report should be a PDF document (created for example from Latex or Word), limited to 5 pages including tables and figures, and may not use fonts less than 10pt in size. You should describe the results of *your analysis* and the conclusions *you* would reach from those results. If you have questions during the analysis, please talk to Kenny or me. You are not allowed to discuss anything related to this assignment with anyone else. This report should look like a formal report to a statistically naive client (i. e. the researcher who brought you the data and/or involved you in the analysis), or an interested lay person. Because a statistical analysis aims to answer a scientific question, you should organize your report in a manner which is customarily used in science. The below gives you some guidance (in large parts, courtesy of Scott Emerson, Department of Biostatistics, University of Washington) on how you could organize your report.

1. **Summary:** Provide a concise description of the question, the data used to try to answer it, and the conclusions of your analysis. Give the most pertinent estimates, confidence intervals, and p-values. Note that estimates and confidence intervals regarding the main question of interest are also important even when there is no statistically significant effect. Don't give too much detail here, but do note any significant problems that were encountered. The basic goal is to have all the key information in your summary, and the rest of your report is the supporting detail.

2. **Background:** Provide a description of the scientific motivation for the analysis. Use your own words rather than copying the description provided by the client. By providing your understanding of the problem, the client may be able to correct any misconceptions that you had about the science. You don't have to go into great detail here, but do give all the facts that entered into your decision process during the analysis.

3. **Questions of Interest:** List the specific questions that your client posed as well as the questions that you answered. Highlight discrepancies between the two categories of questions.

4. **Source of the Data:** Describe the source and sampling methods for the data, if known. Describe the variables that are available and their meaning for the analysis. Highlight patterns of missing data as well as possible confounding by measured or unmeasured variables, if applicable. This should not be a detailed presentation of descriptive statistics, however. That will come under Results.

5. **Statistical Methods:** Describe the methods used for the analysis at two levels. 1) Give a low-level technical description of the analysis for the client to use in the manuscript. Include references for non-standard techniques. You may want to describe the software used, and certainly want to describe the methods used for assessing the appropriateness of your models. Explain how you handled common problems like missing data, multiple comparisons, etc. 2) Explain the basic philosophy behind the analysis techniques in layman's terms. Provide interpretations for all parameter estimates. Motivate transformations. Describe the use of p-values and confidence intervals if they play an important role in your analysis. Explain why you did not use more common techniques if necessary.

6. **Results:** Provide the pertinent results of your analyses. Do not include all the dead-end analyses you might have done unless they provide insight into the question. Do lead the client up to the analyses gradually.

   (a) Start off with descriptive statistics. The goal is to describe the basic characteristics of the sample used to address the question, as well as to present simple descriptive statistics (non-model based) that address the questions. Tables and plots are the key tools. If there are any characteristics of the data that present technical problems that needed to be addressed in the modeling, try to present descriptive statistics illustrating those issues. The basic idea is to presage all the issues you will talk about when presenting the models used in statistical inference, insofar as possible with simple descriptive statistics.

(b) Then go to the major models used to answer the primary questions. Present summaries of the statistical inference obtained from these models (point estimates, confidence intervals, p-values). Highlight any particular issues that materially affected the models used to answer the question (confounding, interactions, nonlinearities, etc.) Tables can often be used to good effect here.

(c) Leave exploratory analyses (if any) for last and highlight the exploratory nature of those analyses. Present the results of your analyses in tables and publishing quality figures. DO NOT INCLUDE OUTPUT FROM STATISTICAL PROGRAMS. Such means little or nothing to a client. When possible, use words instead of cryptic variable names. Use forms of estimates that have some meaning to a statistically naive researcher. Thus, if you log transform your response, present median ratios rather than linear regression parameters. Present confidence intervals rather than the values of Z, t, F, or χ^2 statistics.

7. **Discussion:** Discuss the conclusions which you feel can be drawn from the analyses. Suggest directions for future studies and analyses. Highlight the limitations of the data and your analyses.

8. **Appendix:** Anything of an overly technical nature should be put in an appendix. You may want to include extensive tables in an appendix instead of the main results section. Better yet, don't include anything of an overly technical nature.

The major theme of the above is to write to the client and the scientific community rather than to a statistician. If you cannot explain your findings in a straightforward manner, then the analysis is of little value to the client. Also, lead your reader to all the proper results. You spent a long time analyzing the data. Now provide a brief tour through the high points of your work. Statistical diagnostics, which take a lot of our time, can most often be summarized in a single sentence ("Similar trends were observed at other time points." or "We found no evidence to suggest that the final model did not fit the data adequately"). You are reporting your major results and impressions of the data. If the client wanted to see every detail, he/she would have to do the analysis himself/herself.

Folate Supplementation in Pregnant Women

Background:

Micronutrient deficiencies represent an enormous public health burden in developing countries, affecting well over 1 billion people, with infants, young children and pregnant and lactating women being most vulnerable to their health consequences. Widespread deficiencies in vitamin A or zinc impair host resistance to infection and increase risk of child, infant, and maternal morbidity or mortality. Iron deficiency is the leading cause of anemia, may impair development in children, and increase risk of obstetric complications. Folate deficiency may be a risk factor for heart disease and cancer and, in women, lead to neural tube defects and increase risk of infant mortality. Iodine deficiency is the world's leading cause of preventable mental retardation.

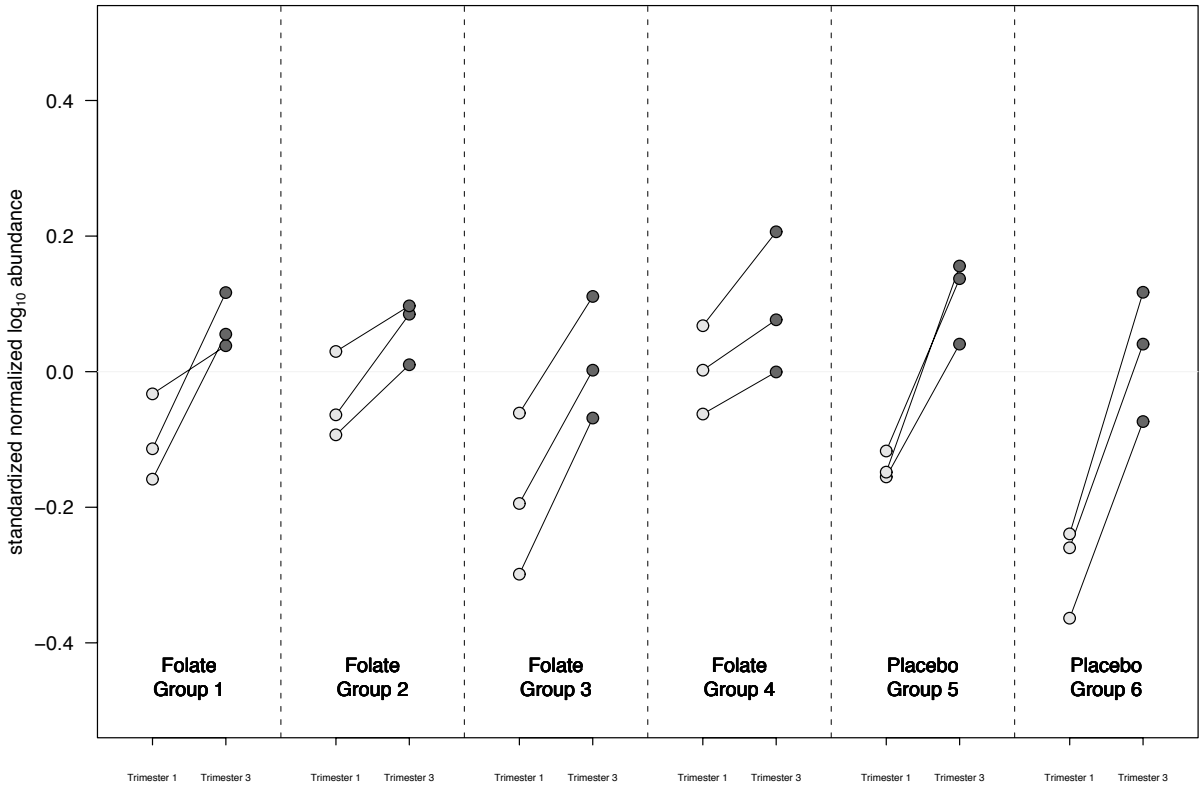
Micronutrients affect gene transcription, protein synthesis, post-translational modification, protein function, and themselves are altered by metabolism mediated by proteins. The plasma represents the most interactive, complete, informative and accessible proteome in the human body. Through measuring, analyzing and identifying hundreds of distinct proteins in minute amounts of specimen by 2-dimensional gel electrophoretic and mass spectrometric techniques, plasma proteomics can be expected to rapidly change the nutritional assessment landscape over the next few years. The application of these new methods can lead to novel indicators of micronutrient status, function, and deficiency, and to the replacement of slow and cumbersome methods to detect micronutrient deficiencies through the development of devices for rapid micronutrient deficiency detection, thereby allowing efficient supplementation of affected subjects.

The Study:

Blood samples from several pregnant women in their first trimester were selected from a population in Nepal. 2D-gel electrophoresis was used to measure protein abundance in the serum (2D-gel electrophoresis is a technique to separate proteins by their mass and charge, and the protein abundance can then be determined by investigating the fluorescent "spots" on the gel). The measure of protein abundance for a particular spot (i. e. , the abundance of a particular protein) is given as standardized normalized $\log_{10}$ abundance, comparing the protein abundance in each of the samples from the pregnant women to the abundance obtained from a background of healthy and non-pregnant women, by taking the

logarithm with base 10 of the abundance ratio. A value of zero for a particular spot obtained from the gel run on the serum of a particular pregnant woman therefore means that the abundance for that protein was equal in that sample to what was expected from the reference. A value of 1 [-1] would mean that the abundance of the protein in the sample was 10 times larger [smaller] that what we would expect from the reference. After the first blood draw, some women were randomly selected and received a folate supplementation, the other women received a placebo. Additional blood samples were obtained in the third trimester of pregnancy for each woman, and gels were run again for each of those samples.

One out of the several hundred protein spots on the gels looked particularly interesting. Data are available for 18 women: 12 received the folate supplementation, and six received a placebo. In both treatment groups, the women were also classified in groups according to some physiological condition, denoted as Group 1 through Group 6 below.



Questions of Interest:

The main goal of the analysis is to determine whether folate supplementation affects the abundance (or here, the standardized normalized $\log_{10}$ abundance, for simplicity referred to as abundance - no transformation of the response necessary) of the protein of interest, and to quantify the variance components. Your analysis should definitely consider (but not be limited to) the following specific questions:

1. Does the folate supplementation affect the abundance of the protein of interest ("treatment effect")?
2. Do the protein abundances differ in the first and the third trimester when no supplementation is administered ("pregnancy effect"), and does that difference depend on the physiological condition recorded?
3. What are the mean protein abundance changes between trimesters in the different groups?
4. Is the between subject variability larger than the variability of what could be considered the experimental error?

Format of the Data:

From `www.biostat.jhsph.edu/~iruczins/teaching/140.752/report/` a file with the data in csv (comma separated value) format can be downloaded, and conveniently read into R using the function `read.csv()`. Here is how the data frame should look like:

	SN.log10.ab	subject	group	treatment	trimester
1	-0.033	1	1	Folate	1
2	0.038	1	1	Folate	3
3	-0.159	2	1	Folate	1
4	0.055	2	1	Folate	3
5	-0.114	3	1	Folate	1
6	0.117	3	1	Folate	3
7	-0.064	4	2	Folate	1
8	0.085	4	2	Folate	3
9	-0.093	5	2	Folate	1
10	0.010	5	2	Folate	3
11	0.030	6	2	Folate	1
12	0.097	6	2	Folate	3
13	-0.061	7	3	Folate	1
14	0.111	7	3	Folate	3
15	-0.299	8	3	Folate	1
16	-0.069	8	3	Folate	3
17	-0.194	9	3	Folate	1
18	0.002	9	3	Folate	3
19	0.002	10	4	Folate	1
20	0.077	10	4	Folate	3
21	0.068	11	4	Folate	1
22	0.206	11	4	Folate	3
23	-0.063	12	4	Folate	1
24	0.000	12	4	Folate	3
25	-0.155	13	5	Placebo	1
26	0.041	13	5	Placebo	3
27	-0.148	14	5	Placebo	1
28	0.156	14	5	Placebo	3
29	-0.117	15	5	Placebo	1
30	0.137	15	5	Placebo	3
31	-0.239	16	6	Placebo	1
32	0.117	16	6	Placebo	3
33	-0.260	17	6	Placebo	1
34	0.041	17	6	Placebo	3
35	-0.364	18	6	Placebo	1
36	-0.074	18	6	Placebo	3