Statistical Issues and Experimental Design in Proteomics

Sreelatha Meleth Ph.D

Format

- Four blocks of 20 minutes
- 10 Minutes for questions
- Small break- depends on number of questions
- ***<u>Have tried to avoid jargon-please ask if anything is not clear</u>***

With respect to Hypothesis Testing & Model Building at the end of this lecture

- You will (should?) know
  - when to use standard statistical tests
    - t-test,, ANOVA, f-test
  - linear / non-linear models
  - Deterministic versus predictive models
  - Univariate versus multivariate techniques
  - Examples of multivariate techniques-PCA, MDS, DA
  - Spectral analysis –
    - Fourier Transforms, Wavelets, Functional Data Analysis

In terms of Statistical Experimental Design

- *You should know*
- What Statistical Experimental Design is:
  Randomization, Replication, Blocking
- Why it adds more to what you do in lab
- What a statistician should know about your experiment and laboratory.
- Experimental design – applied to proteomics

Testing Means (Medians?)

- Mean (t-test) versus median (Wilcoxon RS)
- Statistics aims to estimate quantities that will give us the maximum information about a group by using summary measures.
- Ideally – summary measure is a good 'average' representation of group
- Mean is, in a symmetric population
- Median better in a skewed population
  - – See next slide

2-sample t-test

- Used to see if the mean values of two groups are the same

Assumptions

- Samples drawn from a normal distribution.
- The random variability in the 2 groups is more or less equal
- Assumptions? – the premise used to derive p-values

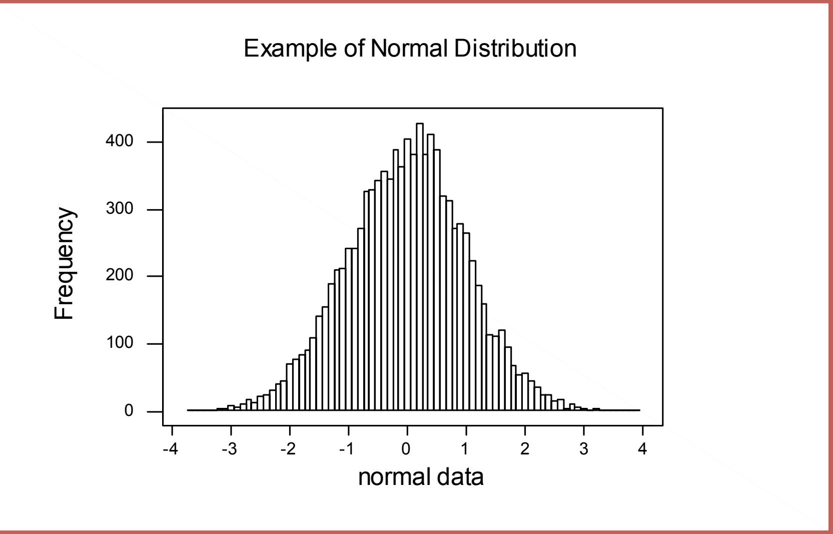
Paired t-test – also used to test means

- Before / After situations
- Non independent observations
- Paired organs- eyes, ears etc
- Uses differences in paired means

T-test cont..

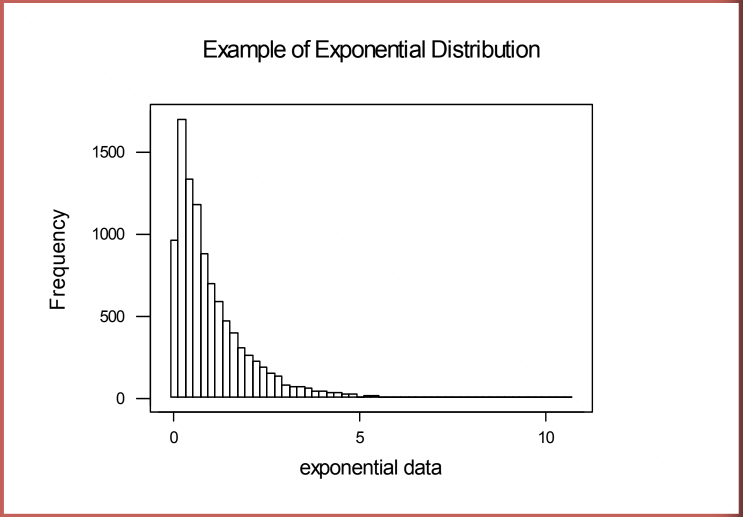
- T-test p-value - what is the probability of observing a difference in means as large as the one observed in the current experiment by chance (fluke occurrence?)
- P-value based on knowledge of t-distribution
- What is a distribution?
- If you plot your measured variable – what does the plot look like?

Pictures of distributions



- Normal Distrbn
- Pr(μ -2σ < Y < μ+2σ)=0.95

- Exponential Dstrbn
- Pr(μ -2σ < Y < μ+2σ)=0.95



Non-parametric (Distribution free) tests?

- When one is not sure of the distribution- really small samples
- Wilcoxon Rank Sum= t-test
- Sign test
- T-test is always more powerful
- Not a bad idea to ask for both tests
- Report only those that are significant in both
- Also available – tests for variance, alternatives for ANOVA & regression

Measuring Change

- Variability- changes in means – currency of statistics
- Comparing the degree of change in means to the degree of change one expects as a matter of course (random error)-basis of ANOVA, Regression models etc
- Is the change caused by the intervention?

F-test

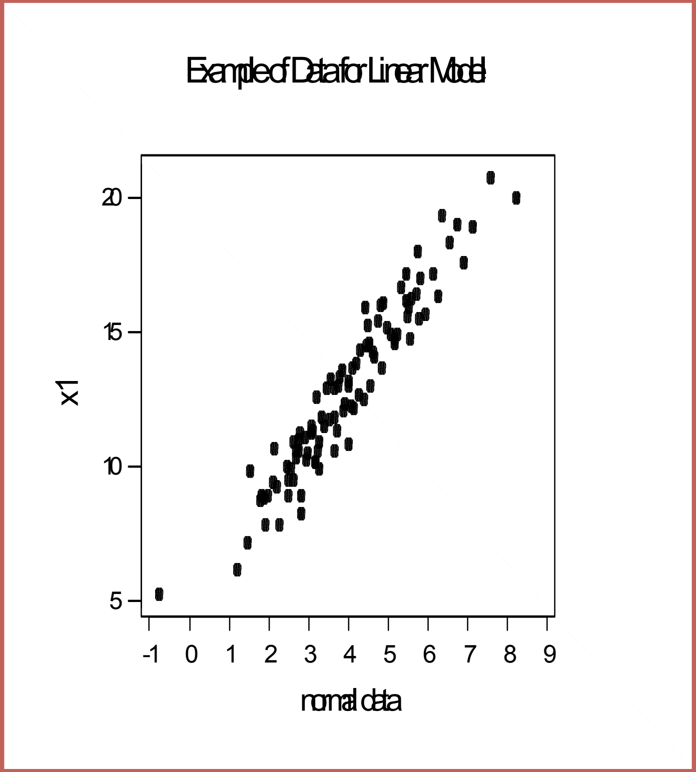
- Compares the variance (degree of change in the means) of the two groups
- Uses the ratio of the variance of one group to the other
- If ratio close to one, the two variances are the same
- If ratio away from one, the two variances are different
- P-value = how likely is it that the size of ratio as high or as small due to chance

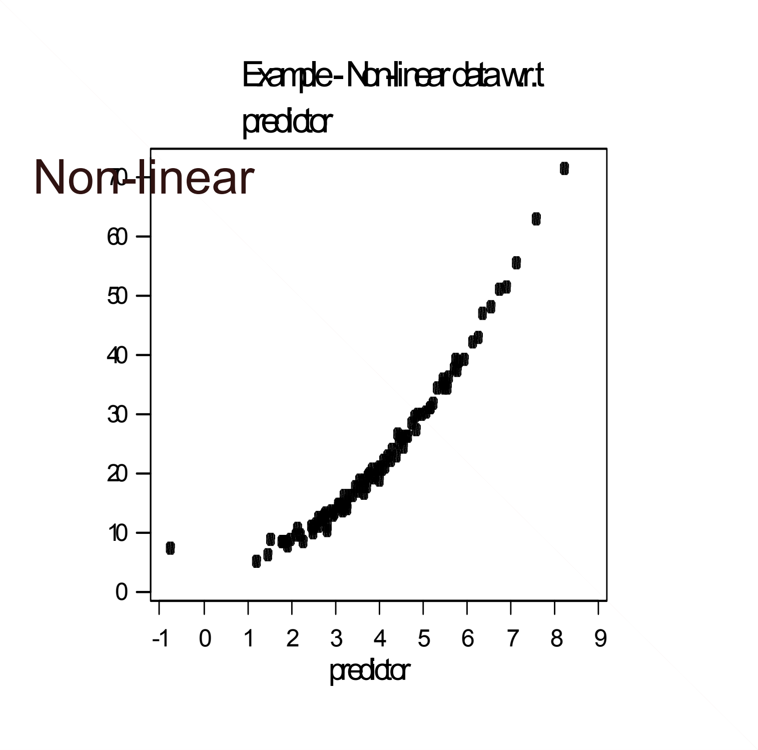
ANOVA

- Extension of t-tests for more than 2 groups
- Statistical experimental design uses extensions of ideas used in ANOVA
- Assumptions identical to those used in t-test
- Although testing means - uses F-tests
- Compares variability caused by factors of interest to random error

Models-linear /non-linear

- Non-linear





Non-linear

Predictive versus deterministic models

- Do you want to find the relationship between variables collected in the current study – e.g. does quantity of reagent affect protein intensity (or has it in the current set of experiments) ?
- Two possibilities once model is built
  - A. To study and understand processes
  - B. Use to design a future experiment with specific quantity of reagent to ensure a certain minimum level of protein intensity

- A is an example of use of model in a deterministic way
- B is same model being used as a predictor

End of first slot… ☺

- Questions?

Univariate methods (e.g.-test, f-test, ANOVA)

For statisticians :

- Univariate = only one outcome variable, e.g. intensity of one protein -WB (?)
- Unit of analysis – row in data set-represents measurement on one individual, e.g., run 20 different WBs – to compare intensity of a particular protein, in a particular tissue in 20 mice fed a particular diet compared to 20 mice fed ordinary diet
- Compare differences intensity in 2 – groups
- A quick check – what test could you use?

Multivariate Methods

- Multivariate = multiple outcomes, e.g., looking simultaneously at intensity of multiple proteins – 2D gels
- The multiple outcomes measured on one unit e.g., multiple intensities on one 2D gel
- Unit of analysis- 2D gel with multiple intensities, the MS spectrum with multiple intensities
- Most techniques aim to use the intensities to group the gels (spectra) into groups
- Groups may be pre-defined (diseased/ non-diseased) and therefore known- supervised classifiers
- Groups may be unknown-is there something that is common to the groups that the proteins fall into – (common function?) – unsupervised classifiers

Brief look at three Multivariate Techniques

- Principal Component Analysis (PCA)
  - Used as a graphical tool
  - Also used as a way of reducing the number of variables in data
  - Exploratory data analysis & visual representations

- Discriminant Analysis (DA)
  - A ‘supervised classifier’ – groups known in advance
  - Almost always used as a predictive model
  - Training data, validation data (sometimes), test data

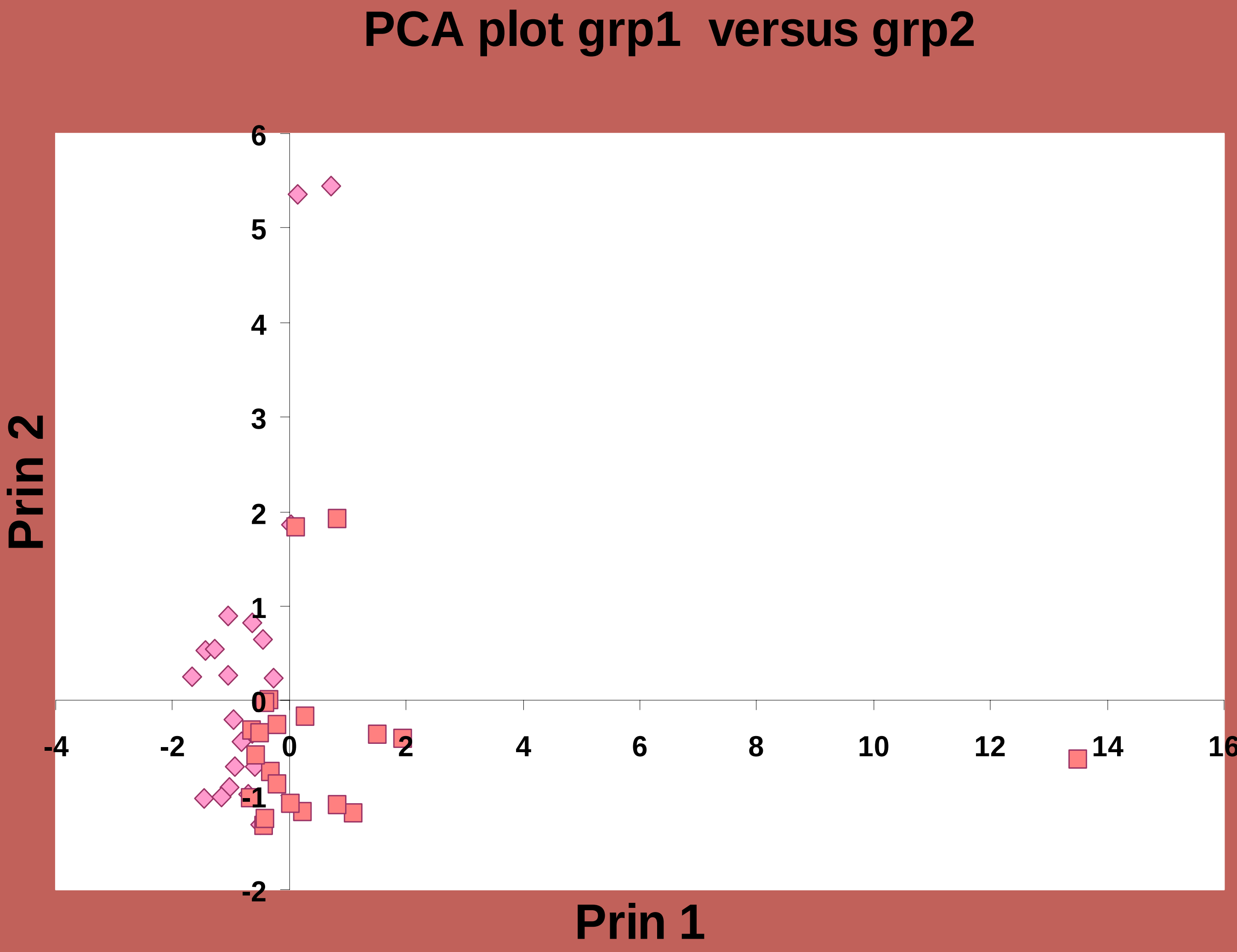
- Multi-Dimensional Scaling
  - Converts dissimilarity (or similarity) measures between variables in a data set to distances
  - Distances used in graphical displays –reveals relationship between groups of proteins – e.g. functional groups of proteins

Principal Components Analysis (PCA)

- Is used in two different ways
  - To help select a small set of variables (e.g. protein intensity) that are good representatives of the entire set of variables
  - A clustering technique-to visualize the relationship between the units of analysis, e.g. do the gels in a particular group cluster together?
- A very brief look at technique
  - Briefly – the technique creates linear combination of multiple outcomes e.g., a + b + c +d - where a, b, c, and d are intensities of individual proteins on a gel (for e.g.)
  - The variances of the combinations are calculated
  - The components (linear combinations) are sorted by their variances – first component has largest, last the least

PCA.. Cont

- If the larger proportion of variability in data is caused by the differences in the groups being studied (e.g., disease/not diseased) then
  - The first few components can be used to represent the variability in the whole data set – reduce dimensions
  - Using pairs of components (e.g. Component 1 versus Component 2) as the two axes in a 2D plot, it is possible to cluster the gels into two groups


PCA plot grp1 versus grp2
Prin 2
6
5
4
3
2
1
0
-1
-2
-4
-2
0
2
4
6
8
10
12
14
16
Prin 1

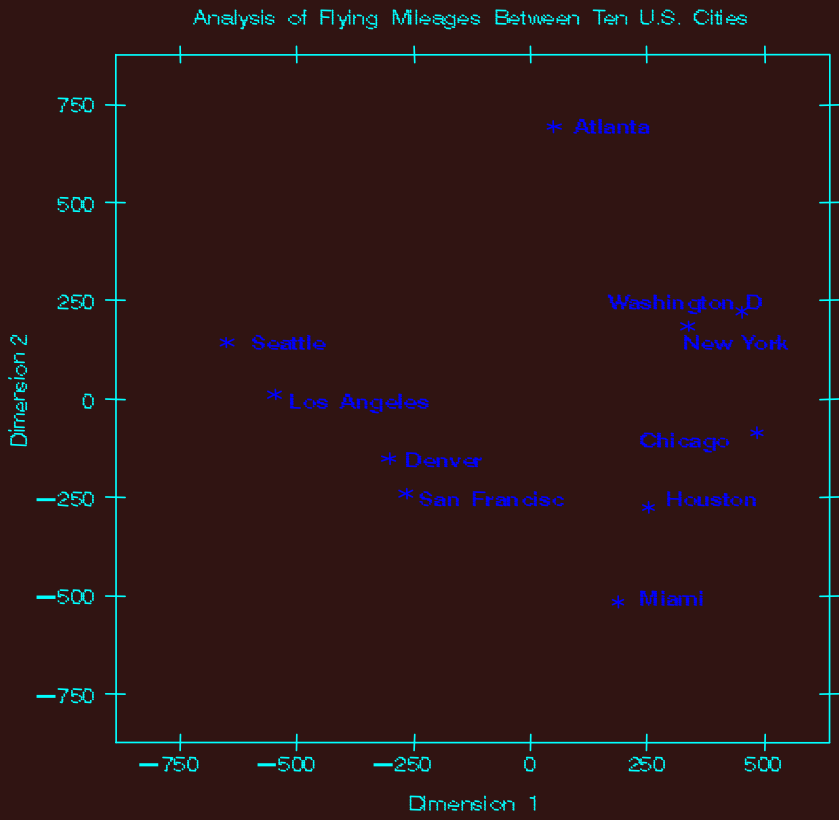
Discriminant Analysis

- Two most common methods – linear and quadratic
  - Linear – outcomes normally distributed, variability in groups more or less the same
  - Quadratic – outcomes normally distributed, variability between groups not necessarily the same
  - Non – parametric methods such as kernel estimation and k-means also available
- The technique uses the product of means and variances – and assigns each unit (gel) to one of the groups based on how close the value of this product is to the product obtained using the mean value for each group

Discriminant Analysis cont..

- The training data is used to select the proteins that will help to classify the groups
- If available the validation data set is used to test the selection, and may be used to re-calibrate, i.e., change the selection of proteins bases on the results of the test
- The test data is used solely to test classifier
- Sensitivity and Specificity of classifier are calculated to measure quality of classifier

Multi Dimensional Scaling



Spectral Analysis

- Reverse process – compared to regression
- Regression – data points – find a line
- In Spectral Analysis – aim to decompose spectra into smaller units
- Fourier Series – uses sin and cos functions to break spectra up –e.g. series of tildes ~ to build the spectra
- Wavelets – similar in concept – different in mathematical function- is said to overcome some of the limitations of Fourier as transforms
- Functional Data Analysis – similar to above

Spectral Analysis.. cont

- In this case the coefficients of the smaller functions are compared with appropriate statistical tests
- The coefficients could be used in classifiers
- Conceptually, -treating the spectra as continuous functions – may not be accurate
- The spectrum is not a ‘true’ spectrum -in that the spectrum is actually a series of distributions of protein ions linked together
- Thus using concepts to measure probability with area under the curve around m_z values seems to do better

End of second section ☺

- Questions?

Statistical Experimental Design

- Measuring variabilty and attributing variability to different sources is a major part of statistical analysis

- Statistical Experimental design – aims to estimate, isolate or neutralize the variabilty

- Uses - Replication, Randomization & Blocking

Replication

- Biological replicates-sample size-power to detect between group
- Technical (same sample) replicates- helps estimate within group variance
- In techniques such as 2D gel, & micro-array technical replicates also help as a quality control measure
- E.g. Are protein spots seen in all replicates of a sample?

Blocking

- Blocking - Create blocks of observations that have very similar variability
- Have every treatment group represented in each group
- e.g., Processing a 2D gel extraneous variability caused by day of processing and / technician involved
- Technicians, day will both be used as blocking factors

Randomization

- After getting a good understanding of process, and variables decide
  - Which variables to block for
  - Which variables are uncontrollable
- Uncontrollable variables neutralized by randomizing across those variables

PI / Statistician interaction

- A number of different designs- CRBD, Latin squares, Split-plots
- Choice depends on close consultation with PI, lab personnel
- Is this design practical?
- You need to say ‘yes it is’, or ‘no it is not’
- Good idea to let statistician to see process in lab

End of third section

- Questions?

Statistical Issues

- Expensive Technologies
- Small samples
- Large number of variables – hi-dimensions
- Lack of experimental design
- Particularly – no replication, no randomization
- Difficult to build good predictive models
- Important to contact statistician in planning stage – not analysis stage